AdaSGN: Adapting Joint Number and Model Size for Efficient Skeleton-Based Action Recognition
Lei Shi, Yifan Zhang, Jian Cheng, Hanqing Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13413-13422
Existing methods for skeleton-based action recognition mainly focus on improving the recognition accuracy, whereas the efficiency of the model is rarely considered. Recently, there are some works trying to speed up the skeleton modeling by designing light-weight modules. However, in addition to the model size, the amount of the data involved in the calculation is also an important factor for the running speed, especially for the skeleton data where most of the joints are redundant or non-informative to identify a specific skeleton.Besides, previous works usually employ one fix-sized model for all the samples regardless of the difficulty of recognition, which wastes computations for easy samples.To address these limitations, a novel approach, called AdaSGN, is proposed in this paper, which can reduce the computational cost of the inference process by adaptively controlling the input number of the joints of the skeleton on-the-fly. Moreover, it can also adaptively select the optimal model size for each sample to achieve a better trade-off between the accuracy and the efficiency. We conduct extensive experiments on three challenging datasets, namely, NTU-60, NTU-120 and SHREC, to verify the superiority of the proposed approach, where AdaSGN achieves comparable or even higher performance with much lower GFLOPs compared with the baseline method.
****

C2N: Practical Generative Noise Modeling for Real-World Denoising
Geonwoon Jang, Wooseok Lee, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2350-2359
Learning-based image denoising methods have been bounded to situations where well-aligned noisy and clean images are given, or samples are synthesized from predetermined noise models, e.g., Gaussian. While recent generative noise modeling methods aim to simulate the unknown distribution of real-world noise, several limitations still exist. In a practical scenario, a noise generator should learn to simulate the general and complex noise distribution without using paired noisy and clean images. However, since existing methods are constructed on the unrealistic assumption of real-world noise, they tend to generate implausible patterns and cannot express complicated noise maps. Therefore, we introduce a Clean-to-Noisy image generation framework, namely C2N, to imitate complex real-world noise without using any paired examples. We construct the noise generator in C2N accordingly with each component of real-world noise characteristics to express a wide range of noise accurately. Combined with our C2N, conventional denoising CNNs can be trained to outperform existing unsupervised methods on challenging real-world benchmarks by a large margin.
****

Continual Learning on Noisy Data Streams via Self-Purified Replay
Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, Gunhee Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 537-547
Continually learning in the real world must overcome many challenges, among which noisy labels are a common and inevitable issue. In this work, we present a replay-based continual learning framework that simultaneously addresses both catastrophic forgetting and noisy labels for the first time. Our solution is based on two observations; (i) forgetting can be mitigated even with noisy labels via self-supervised learning, and (ii) the purity of the replay buffer is crucial. Building on this regard, we propose two key components of our method: (i) a self-supervised replay technique named Self-Replay, which can circumvent erroneous training signals arising from noisy labeled data, and (ii) the Self-Centered filter that maintains a purified replay buffer via centrality-based stochastic graph ensembles. The empirical results on MNIST, CIFAR-10, CIFAR-100, and WebVision with real-world noise demonstrate that our framework can maintain a highly pure replay buffer amidst noisy streamed data while greatly outperforming the combinations of the state-of-the-art continual learning and noisy label learning methods.
****

FOVEA: Foveated Image Magnification for Autonomous Navigation
Chittesh Thavamani, Mengtian Li, Nicolas Cebron, Deva Ramanan; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15539-1 5548
Efficient processing of high-resolution video streams is safety-critical for man y robotics applications such as autonomous driving. Image downsampling is a comm only adopted technique to ensure the latency constraint is met. However, this na ive approach greatly restricts an object detector's capability to identify small objects. In this paper, we propose an attentional approach that elastically mag nifies certain regions while maintaining a small input canvas. The magnified reg ions are those that are believed to have a high probability of containing an obj ect, whose signal can come from a dataset-wide prior or frame-level prior comput ed from recent object predictions. The magnification is implemented by a KDE-bas ed mapping to transform the bounding boxes into warping parameters, which are th en fed into an image sampler with anti-cropping regularization. The detector is then fed with the warped image and we apply a differentiable backward mapping to get bounding box outputs in the original space. Our regional magnification allo ws algorithms to make better use of high-resolution input without incurring the cost of high-resolution processing. On the autonomous driving datasets Argoverse -HD and BDD100K, we show our proposed method boosts the detection AP over standa rd Faster R-CNN, with and without finetuning. Additionally, building on top of t he previous state-of-the-art in streaming detection, our method sets a new recor d for streaming AP on Argoverse-HD (from 17.8 to 23.0 on a GTX 1080 Ti GPU), sug gesting that it has achieved a superior accuracy-latency tradeoff.
****

PlenOctrees for Real-Time Rendering of Neural Radiance Fields
Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, Angjoo Kanazawa; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5752-5761
We introduce a method to render Neural Radiance Fields (NeRFs) in real time usin g PlenOctrees, an octree-based 3D representation which supports view-dependent e ffects. Our method can render 800x800 images at more than 150 FPS, which is over 3000 times faster than conventional NeRFs. We do so without sacrificing quality while preserving the ability of NeRFs to perform free-viewpoint rendering of sc enes with arbitrary geometry and view-dependent effects. Real-time performance i s achieved by pre-tabulating the NeRF into a PlenOctree. In order to preserve vi ew-dependent effects such as specularities, we factorize the appearance via clos ed-form spherical basis functions. Specifically, we show that it is possible to train NeRFs to predict a spherical harmonic representation of radiance, removing the viewing direction as an input to the neural network. Furthermore, we show t hat PlenOctrees can be directly optimized to further minimize the reconstruction loss, which leads to equal or better quality compared to competing methods. Mor eover, this octree optimization step can be used to reduce the training time, as we no longer need to wait for the NeRF training to converge fully. Our real-tim e neural rendering approach may potentially enable new applications such as 6-DO F industrial and product visualizations, as well as next generation AR/VR system s. PlenOctrees are amenable to in-browser rendering as well; please visit the pr oject page for the interactive online demo, as well as video and code: https://a lexyu.net/plenoctrees.
****

Entropy Maximization and Meta Classification for Out-of-Distribution Detection i n Semantic Segmentation
Robin Chan, Matthias Rottmann, Hanno Gottschalk; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2021, pp. 5128-5137
Deep neural networks (DNNs) for the semantic segmentation of images are usually trained to operate on a predefined closed set of object classes. This is in cont rast to the ""open world"" setting where DNNs are envisioned to be deployed to. From a functional safety point of view, the ability to detect so-called ""out-of -distribution"" (OoD) samples, i.e., objects outside of a DNN's semantic space, is crucial for many applications such as automated driving. A natural baseline a

pproach to OoD detection is to threshold on the pixel-wise softmax entropy. We p resent a two-step procedure that significantly improves that approach. Firstly, we utilize samples from the COCO dataset as OoD proxy and introduce a second tra ining objective to maximize the softmax entropy on these samples. Starting from pretrained semantic segmentation networks we re-train a number of DNNs on differ ent in-distribution datasets and consistently observe improved OoD detection per formance when evaluating on completely disjoint OoD datasets. Secondly, we perfo rm a transparent post-processing step to discard false positive OoD samples by s o-called ""meta classification"". To this end, we apply linear models to a set o f hand-crafted metrics derived from the DNN's softmax probabilities. In our expe riments we consistently observe a clear additional gain in OoD detection perform ance, cutting down the number of detection errors by 52% when comparing the best baseline with our results. We achieve this improvement sacrificing only margina lly in original segmentation performance. Therefore, our method contributes to s afer DNNs with more reliable overall system performance.
****

Specificity-Preserving RGB-D Saliency Detection
Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, Ling Shao; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 468 1-4691
RGB-D saliency detection has attracted increasing attention, due to its effectiv eness and the fact that depth cues can now be conveniently captured. Existing wo rks often focus on learning a shared representation through various fusion strat egies, with few methods explicitly considering how to preserve modality-specific characteristics. In this paper, taking a new perspective, we propose a specific ity-preserving network for RGB-D saliency detection, which benefits saliency det ection performance by exploring both the shared information and modality-specifi c properties (e.g., specificity). Specifically, two modality-specific networks a nd a shared learning network are adopted to generate individual and shared salie ncy maps. A cross-enhanced integration module (CIM) is proposed to fuse cross-mo dal features in the shared learning network, which are then propagated to the ne xt layer for integrating cross-level information. Besides, we propose a multi-mo dal feature aggregation (MFA) module to integrate the modality-specific features from each individual decoder into the shared decoder, which can provide rich co mplementary multi-modal information to boost the saliency detection performance. Further, a skip connection is used to combine hierarchical features between the encoder and decoder layers. Experiments on six benchmark datasets demonstrate t hat our SP-Net outperforms other state-of-the-art methods.
****

3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds
Lichen Zhao, Daigang Cai, Lu Sheng, Dong Xu; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 2928-2937
Visual grounding on 3D point clouds is an emerging vision and language task that benefits various applications in understanding the 3D visual world. By formulat ing this task as a grounding-by-detection problem, lots of recent works focus on how to exploit more powerful detectors and comprehensive language features, but (1) how to model complex relations for generating context-aware object proposal s and (2) how to leverage proposal relations to distinguish the true target obje ct from similar proposals are not fully studied yet. Inspired by the well-known transformer architecture, we propose a relation-aware visual grounding method on 3D point clouds, named as 3DVG-Transformer, to fully utilize the contextual clu es for relationenhanced proposal generation and cross-modal proposal disambiguat ion, which are enabled by a newly designed coordinate-guided contextual aggregat ion (CCA) module in the object proposal generation stage, and a multiplex attent ion (MA) module in the cross-modal feature fusion stage. We validate that our 3D VG-Transformer outperforms the state-of-the-art methods by a large margin, on tw o point cloud-based visual grounding datasets, ScanRefer and Nr3D/Sr3D from Refe rIt3D, especially for complex scenarios containing multiple objects of the same category.
****

4D-Net for Learned Multi-Modal Alignment
AJ Piergiovanni, Vincent Casser, Michael S. Ryoo, Anelia Angelova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15435-15445
We present 4D-Net, a 3D object detection approach, which utilizes 3D Point Cloud and RGB sensing information, both in time. We are able to incorporate the 4D information by performing a novel dynamic connection learning across various feature representations and levels of abstraction and by observing geometric constraints. Our approach outperforms the state-of-the-art and strong baselines on the Waymo Open Dataset. 4D-Net is better able to use motion cues and dense image information to detect distant objects more successfully. We will open source the code.
****

Patch Craft: Video Denoising by Deep Modeling and Patch Matching
Gregory Vaksman, Michael Elad, Peyman Milanfar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2157-2166
The non-local self-similarity property of natural images has been exploited extensively for solving various image processing problems. When it comes to video sequences, harnessing this force is even more beneficial due to the temporal redundancy. In the context of image and video denoising, many classically-oriented algorithms employ self-similarity, splitting the data into overlapping patches, gathering groups of similar ones and processing these together somehow. With the emergence of convolutional neural networks (CNN), the patch-based framework has been abandoned. Most CNN denoisers operate on the whole image, leveraging non-local relations only implicitly by using a large receptive field. This work proposes a novel approach for leveraging self-similarity in the context of video denoising, while still relying on a regular convolutional architecture. We introduce a concept of patch-craft frames - artificial frames that are similar to the real ones, built by tiling matched patches. Our algorithm augments video sequences with patch-craft frames and feeds them to a CNN. We demonstrate the substantial boost in denoising performance obtained with the proposed approach.
****

Image Manipulation Detection by Multi-View Multi-Scale Supervision
Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, Xirong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14185-14193
The key challenge of image manipulation detection is how to learn generalizable features that are sensitive to manipulations in novel data, whilst specific to prevent false alarms on authentic images. Current research emphasizes the sensitivity, with the specificity overlooked. In this paper we address both aspects by multi-view feature learning and multi-scale supervision. By exploiting noise distribution and boundary artifact surrounding tampered regions, the former aims to learn semantic-agnostic and thus more generalizable features. The latter allows us to learn from authentic images which are nontrivial to taken into account by current semantic segmentation network based methods. Our thoughts are realized by a new network which we term MVSS-Net. Extensive experiments on five benchmark sets justify the viability of MVSS-Net for both pixel-level and image-level manipulation detection.
****

Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation
Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, Cuihua Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15520-15528
Large-scale point cloud semantic segmentation has wide applications. Current popular researches mainly focus on fully supervised learning which demands expensive and tedious manual point-wise annotation. Weakly supervised learning is an alternative way to avoid this exhausting annotation. However, for large-scale point clouds with few labeled points, the network is difficult to extract discriminative features for unlabeled points, as well as the regularization of topology between labeled and unlabeled points is usually ignored, resulting in incorrect seg

mentation results. To address this problem, we propose a perturbed self-distillation (PSD) framework. Specifically, inspired by self-supervised learning, we construct the perturbed branch and enforce the predictive consistency among the perturbed branch and original branch. In this way, the graph topology of the whole point cloud can be effectively established by the introduced auxiliary supervision, such that the information propagation between the labeled and unlabeled points will be realized. Besides point-level supervision, we present a well-integrated context-aware module to explicitly regularize the affinity correlation of labeled points. Therefore, the graph topology of the point cloud can be further refined. The experimental results evaluated on three large-scale datasets show the large gain (3.0% on average) against recent weakly supervised methods and comparable results to some fully supervised methods.

****

Cherry-Picking Gradients: Learning Low-Rank Embeddings of Visual Data via Differentiable Cross-Approximation

Mikhail Usvyatsov, Anastasia Makarova, Rafael Ballester-Ripoll, Maxim Rakhuba, Andreas Krause, Konrad Schindler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11426-11435

We propose an end-to-end trainable framework that processes large-scale visual data tensors by looking at a fraction of their entries only. Our method combines a neural network encoder with a tensor train decomposition to learn a low-rank latent encoding, coupled with cross-approximation (CA) to learn the representation through a subset of the original samples. CA is an adaptive sampling algorithm that is native to tensor decompositions and avoids working with the full high-resolution data explicitly. Instead, it actively selects local representative samples that we fetch out-of-core and on demand. The required number of samples grows only logarithmically with the size of the input. Our implicit representation of the tensor in the network enables processing large grids that could not be otherwise tractable in their uncompressed form. The proposed approach is particularly useful for large-scale multidimensional grid data (e.g., 3D tomography), and for tasks that require context over a large receptive field (e.g., predicting the medical condition of entire organs). The code is available at https://github.com/aelphy/c-pic.

****

Ask&Confirm: Active Detail Enriching for Cross-Modal Retrieval With Partial Query

Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, Xing Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1835-1844

Text-based image retrieval has seen considerable progress in recent years. However, the performance of existing methods suffers in real life since the user is likely to provide an incomplete description of an image, which often leads to results filled with false positives that fit the incomplete description. In this work, we introduce the partial-query problem and extensively analyze its influence on text-based image retrieval. Previous interactive methods tackle the problem by passively receiving users' feedback to supplement the incomplete query iteratively, which is time-consuming and requires heavy user effort. Instead, we propose a novel retrieval framework that conducts the interactive process in an Ask-and-Confirm fashion, where AI actively searches for discriminative details missing in the current query, and users only need to confirm AI's proposal. Specifically, we propose an object-based interaction to make the interactive retrieval more user-friendly and present a reinforcement-learning-based policy to search for discriminative objects. Furthermore, since fully-supervised training is often infeasible due to the difficulty of obtaining human-machine dialog data, we present a weakly-supervised training strategy that needs no human-annotated dialogs other than a text-image dataset. Experiments show that our framework significantly improves the performance of text-based image retrieval. Code is available at https://github.com/CuthbertCai/Ask-Confirm.

****

EventHands: Real-Time Neural 3D Hand Pose Estimation From an Event Stream

Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, Christian Theobalt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12385-12395
3D hand pose estimation from monocular videos is a long-standing and challenging problem, which is now seeing a strong upturn. In this work, we address it for the first time using a single event camera, i.e., an asynchronous vision sensor reacting on brightness changes. Our EventHands approach has characteristics previously not demonstrated with a single RGB or depth camera such as high temporal resolution at low data throughputs and real-time performance at 1000 Hz. Due to the different data modality of event cameras compared to classical cameras, existing methods cannot be directly applied to and re-trained for event streams. We thus design a new neural approach which accepts a new event stream representation suitable for learning, which is trained on newly-generated synthetic event streams and can generalise to real data. Experiments show that EventHands outperforms recent monocular methods using a colour (or depth) camera in terms of accuracy and its ability to capture hand motions of unprecedented speed. Our method, the event stream simulator and the dataset are publicly available (see https://gvv.mpi-inf.mpg.de/projects/EventHands/).
****

Composable Augmentation Encoding for Video Representation Learning
Chen Sun, Arsha Nagrani, Yonglong Tian, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8834-8844
We focus on contrastive methods for self-supervised video representation learning. A common paradigm in contrastive learning is to construct positive pairs by sampling different data views for the same instance, with different data instances as negatives. These methods implicitly assume a set of representational invariances to the view selection mechanism (e.g., sampling frames with temporal shifts), which may lead to poor performance on downstream tasks which violate these invariances (fine-grained video action recognition that would benefit from temporal information). To overcome this limitation, we propose an `augmentation aware' contrastive learning framework, where we explicitly provide a sequence of augmentation parameterisations (such as the values of the time shifts used to create data views) as composable augmentation encodings (CATE) to our model when projecting the video representations for contrastive learning. We show that representations learned by our method encode valuable information about specified spatial or temporal augmentation, and in doing so also achieve state-of-the-art performance on a number of video benchmarks.
****

Exploiting Explanations for Model Inversion Attacks
Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, Brian Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 682-692
The successful deployment of artificial intelligence (AI) in many domains from healthcare to hiring requires their responsible use, particularly in model explanations and privacy. Explainable artificial intelligence (XAI) provides more information to help users to understand model decisions, yet this additional knowledge exposes additional risks for privacy attacks. Hence, providing explanation harms privacy. We study this risk for image-based model inversion attacks and identified several attack architectures with increasing performance to reconstruct private image data from model explanations. We have developed several multi-modal transposed CNN architectures that achieve significantly higher inversion performance than using the target model prediction only. These XAI-aware inversion models were designed to exploit the spatial knowledge in image explanations. To understand which explanations have higher privacy risk, we analyzed how various explanation types and factors influence inversion performance. In spite of some models not providing explanations, we further demonstrate increased inversion performance even for non-explainable target models by exploiting explanations of surrogate models through attention transfer. This method first inverts an explanation from the target prediction, then reconstructs the target image. These threats highlight the urgent and significant privacy risks of explanations and calls attention for new privacy preservation techniques that balance the dual-requirement

for AI explainability and privacy.
****

Semantic Diversity Learning for Zero-Shot Multi-Label Classification
Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 640-650
Training a neural network model for recognizing multiple labels associated with an image, including identifying unseen labels, is challenging, especially for images that portray numerous semantically diverse labels. As challenging as this task is, it is an essential task to tackle since it represents many real-world cases, such as image retrieval of natural images. We argue that using a single embedding vector to represent an image, as commonly practiced, is not sufficient to rank both relevant seen and unseen labels accurately. This study introduces an end-to-end model training for multi-label zero-shot learning that supports the semantic diversity of the images and labels. We propose to use an embedding matrix having principal embedding vectors trained using a tailored loss function. In addition, during training, we suggest up-weighting in the loss function image samples presenting higher semantic diversity to encourage the diversity of the embedding matrix. Extensive experiments show that our proposed method improves the zero-shot model's quality in tag-based image retrieval achieving SoTA results on several common datasets (NUS-Wide, COCO, Open Images).
****

Describing and Localizing Multiple Changes With Transformers
Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, Yutaka Satoh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1971-1980
Existing change captioning studies have mainly focused on a single change. However, detecting and describing multiple changed parts in image pairs is essential for enhancing adaptability to complex scenarios. We solve the above issues from three aspects: (i) We propose a simulation-based multi-change captioning dataset; (ii) We benchmark existing state-of-the-art methods of single change captioning on multi-change captioning; (iii) We further propose Multi-Change Captioning transformers (MCCFormers) that identify change regions by densely correlating different regions in image pairs and dynamically determines the related change regions with words in sentences. The proposed method obtained the highest scores on four conventional change captioning evaluation metrics for multi-change captioning. Additionally, our proposed method can separate attention maps for each change and performs well with respect to change localization. Moreover, the proposed framework outperformed the previous state-of-the-art methods on an existing change captioning benchmark, CLEVR-Change, by a large margin (+6.1 on BLEU-4 and +9.7 on CIDEr scores), indicating its general ability in change captioning tasks. The code and dataset are available at the project page.
****

Score-Based Point Cloud Denoising
Shitong Luo, Wei Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4583-4592
Point clouds acquired from scanning devices are often perturbed by noise, which affects downstream tasks such as surface reconstruction and analysis. The distribution of a noisy point cloud can be viewed as the distribution of a set of noise-free samples p(x) convolved with some noise model n, leading to (p * n)(x) whose mode is the underlying clean surface. To denoise a noisy point cloud, we propose to increase the log-likelihood of each point from p * n via gradient ascent---iteratively updating each point's position. Since p * n is unknown at test-time, and we only need the score (i.e., the gradient of the log-probability function) to perform gradient ascent, we propose a neural network architecture to estimate the score of p * n given only noisy point clouds as input. We derive objective functions for training the network and develop a denoising algorithm leveraging on the estimated scores. Experiments demonstrate that the proposed model outperforms state-of-the-art methods under a variety of noise models, and shows the potential to be applied in other tasks such as point cloud upsampling.

****
Panoptic Segmentation of Satellite Image Time Series With Convolutional Temporal Attention Networks
Vivien Sainte Fare Garnot, Loic Landrieu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4872-4881
Unprecedented access to multi-temporal satellite imagery has opened new perspectives for a variety of Earth observation tasks. Among them, pixel-precise panoptic segmentation of agricultural parcels has major economic and environmental implications. While researchers have explored this problem for single images, we argue that the complex temporal patterns of crop phenology are better addressed with temporal sequences of images. In this paper, we present the first end-to-end, single-stage method for panoptic segmentation of Satellite Image Time Series (SITS). This module can be combined with our novel image sequence encoding network which relies on temporal self-attention to extract rich and adaptive multi-scale spatio-temporal features. We also introduce PASTIS, the first open-access SITS dataset with panoptic annotations. We demonstrate the superiority of our encoder for semantic segmentation against multiple competing network architectures, and set up the first state-of-the-art of panoptic segmentation of SITS. Our implementation and the PASTIS dataset are publicly available at (link-upon-publication).
****
Focus on the Positives: Self-Supervised Learning for Biodiversity Monitoring
Omiros Pantazis, Gabriel J. Brostow, Kate E. Jones, Oisin Mac Aodha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10583-10592
We address the problem of learning self-supervised representations from unlabeled image collections. Unlike existing approaches that attempt to learn useful features by maximizing similarity between augmented versions of each input image or by speculatively picking negative samples, we instead also make use of the natural variation that occurs in image collections that are captured using static monitoring cameras. To achieve this, we exploit readily available context data that encodes information such as the spatial and temporal relationships between the input images. We are able to learn representations that are surprisingly effective for downstream supervised classification, by first identifying high probability positive pairs at training time, i.e. those images that are likely to depict the same visual concept. For the critical task of global biodiversity monitoring, this results in image features that can be adapted to challenging visual species classification tasks with limited human supervision. We present results on four different camera trap image collections, across three different families of self-supervised learning methods, and show that careful image selection at training time results in superior performance compared to existing baselines such as conventional self-supervised training and transfer learning.
****
Bridging Unsupervised and Supervised Depth From Focus via All-in-Focus Supervision
Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, Kevin Jou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12621-12631
Depth estimation is a long-lasting yet important task in computer vision. Most of the previous works try to estimate depth from input images and assume images are all-in-focus (AiF), which is less common in real-world applications. On the other hand, a few works take defocus blur into account and consider it as another cue for depth estimation. In this paper, we propose a method to estimate not only a depth map but an AiF image from a set of images with different focus positions (known as a focal stack). We design a shared architecture to exploit the relationship between depth and AiF estimation. As a result, the proposed method can be trained either supervisedly with ground truth depth, or unsupervisedly with AiF images as supervisory signals. We show in various experiments that our method outperforms the state-of-the-art methods both quantitatively and qualitatively, and also has higher efficiency in inference time.

****
Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions
Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 568-578
Although convolutional neural networks (CNNs) have achieved great success in computer vision, this work investigates a simpler, convolution-free backbone network useful for many dense prediction tasks. Unlike the recently-proposed Vision Transformer (ViT) that was designed for image classification specifically, we introduce the Pyramid Vision Transformer (PVT), which overcomes the difficulties of porting Transformer to various dense prediction tasks. PVT has several merits compared to current state of the arts. (1) Different from ViT that typically yields low-resolution outputs and incurs high computational and memory costs, PVT not only can be trained on dense partitions of an image to achieve high output resolution, which is important for dense prediction, but also uses a progressive shrinking pyramid to reduce the computations of large feature maps. (2) PVT inherits the advantages of both CNN and Transformer, making it a unified backbone for various vision tasks without convolutions, where it can be used as a direct replacement for CNN backbones. (3) We validate PVT through extensive experiments, showing that it boosts the performance of many downstream tasks, including object detection, instance and semantic segmentation. For example, with a comparable number of parameters, PVT+RetinaNet achieves 40.4 AP on the COCO dataset, surpassing ResNet50+RetinNet (36.3 AP) by 4.1 absolute AP. We hope that PVT could serve as an alternative and useful backbone for pixel-level predictions and facilitate future research.
****
DOLG: Single-Stage Image Retrieval With Deep Orthogonal Fusion of Local and Global Features
Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, Jizhou Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11772-11781
Image Retrieval is a fundamental task of obtaining images similar to the query one from a database. A common image retrieval practice is to firstly retrieve candidate images via similarity search using global image features and then re-rank the candidates by leveraging their local features. Previous learning-based studies mainly focus on either global or local image representation learning to tackle the retrieval task. In this paper, we abandon the two-stage paradigm and seek to design an effective single-stage solution by integrating local and global information inside images into compact image representations. Specifically, we propose a Deep Orthogonal Local and Global (DOLG) information fusion framework for end-to-end image retrieval. It attentively extracts representative local information with multi-atrous convolutions and self-attention at first. Components orthogonal to the global image representation are then extracted from the local information. At last, the orthogonal components are concatenated with the global representation as a complementary, and then aggregation is performed to generate the final representation. The whole framework is end-to-end differentiable and can be trained with image-level labels. Extensive experimental results validate the effectiveness of our solution and show that our model achieves state-of-the-art image retrieval performances on Revisited Oxford and Paris datasets.
****
Light Source Guided Single-Image Flare Removal From Unpaired Data
Xiaotian Qiao, Gerhard P. Hancke, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4177-4185
Causally-taken images often suffer from flare artifacts, due to the unintended reflections and scattering of light inside the camera. However, as flares may appear in a variety of shapes, positions, and colors, detecting and removing them entirely from an image is very challenging. Existing methods rely on predefined intensity and geometry priors of flares, and may fail to distinguish the difference between light sources and flare artifacts. We observe that the conditions of

the light source in the image play an important role in the resulting flares. In this paper, we present a deep framework with light source aware guidance for single-image flare removal (SIFR). In particular, we first detect the light source regions and the flare regions separately, and then remove the flare artifacts based on the light source aware guidance. By learning the underlying relationships between the two types of regions, our approach can remove different kinds of flares from the image. In addition, instead of using paired training data which are difficult to collect, we propose the first unpaired flare removal dataset and new cycle-consistency constraints to obtain more diverse examples and avoid manual annotations. Extensive experiments demonstrate that our method outperforms the baselines qualitatively and quantitatively. We also show that our model can be applied to flare effect manipulation (e.g., adding or changing image flares).
****

Learning Bias-Invariant Representation by Cross-Sample Mutual Information Minimization
Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, Jiebo Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15002-15012
Deep learning algorithms mine knowledge from the training data and thus would likely inherit the dataset's bias information. As a result, the obtained model would generalize poorly and even mislead the decision process in real-life applications. We propose to remove the bias information misused by the target task with a cross-sample adversarial debiasing (CSAD) method. CSAD explicitly extracts target and bias features disentangled from the latent representation generated by a feature extractor and then learns to discover and remove the correlation between the target and bias features. The correlation measurement plays a critical role in adversarial debiasing and is conducted by a cross-sample neural mutual information estimator. Moreover, we propose joint content and local structural representation learning to boost mutual information estimation for better performance. We conduct thorough experiments on publicly available datasets to validate the advantages of the proposed method over state-of-the-art approaches.
****

Selective Feature Compression for Efficient Activity Recognition Inference
Chunhui Liu, Xinyu Li, Hao Chen, Davide Modolo, Joseph Tighe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13628-13637
Most action recognition solutions rely on dense sampling to precisely cover the informative temporal clip. Extensively searching temporal region is expensive for a real-world application. In this work, we focus on improving the inference efficiency of current action recognition backbones on trimmed videos, and illustrate that one action model can also cover then informative region by dropping non-informative features. We present Selective Feature Compression (SFC), an action recognition inference strategy that greatly increase model inference efficiency without any accuracy compromise. Differently from previous works that compress kernel sizes and decrease the channel dimension, we propose to compress feature flow at spatio-temporal dimension without changing any backbone parameters. Our experiments on Kinetics-400, UCF101 and ActivityNet show that SFC is able to reduce inference speed by 6-7x and memory usage by 5-6x compared with the commonly used 30 crops dense sampling procedure, while also slightly improving Top1 Accuracy. We thoroughly quantitatively and qualitatively evaluate SFC and all its components and show how does SFC learn to attend to important video regions and to drop temporal features that are uninformative for the task of action recognition.
****

Attention-Based Multi-Reference Learning for Image Super-Resolution
Marco Pesavento, Marco Volino, Adrian Hilton; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14697-14706
This paper proposes a novel Attention-based Multi-Reference Super-resolution network (AMRSR) that, given a low-resolution image, learns to adaptively transfer the most similar texture from multiple reference images to the super-resolution output whilst maintaining spatial coherence. The use of multiple reference images together with attention-based sampling is demonstrated to achieve significantly

improved performance over state-of-the-art reference super-resolution approaches on multiple benchmark datasets. Reference super-resolution approaches have recently been proposed to overcome the ill-posed problem of image super-resolution by providing additional information from a high-resolution reference image. Multi-reference super-resolution extends this approach by providing a more diverse pool of image features to overcome the inherent information deficit whilst maintaining memory efficiency. A novel hierarchical attention-based sampling approach is introduced to learn the similarity between low-resolution image features and multiple reference images based on a perceptual loss. Ablation demonstrates the contribution of both multi-reference and hierarchical attention-based sampling to overall performance. Perceptual and quantitative ground-truth evaluation demonstrates significant improvement in performance even when the reference images deviate significantly from the target image. The project website can be found at https://marcopesavento.github.io/AMRSR/

****

Spatial-Temporal Transformer for Dynamic Scene Graph Generation

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, Michael Ying Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16372-16382

Dynamic scene graph generation aims at generating a scene graph of the given video. Compared to the task of scene graph generation from images, it is more challenging because of the dynamic relationships between objects and the temporal dependencies between frames allowing for a richer semantic interpretation. In this paper, we propose Spatial-temporal Transformer (STTran), a neural network that consists of two core modules: (1) a spatial encoder that takes an input frame to extract spatial context and reason about the visual relationships within a frame, and (2) a temporal decoder which takes the output of the spatial encoder as input in order to capture the temporal dependencies between frames and infer the dynamic relationships. Furthermore, STTran is flexible to take varying lengths of videos as input without clipping, which is especially important for long videos. Our method is validated on the benchmark dataset Action Genome (AG). The experimental results demonstrate the superior performance of our method in terms of dynamic scene graphs. Moreover, a set of ablative studies is conducted and the effect of each proposed module is justified.

****

Deep Transport Network for Unsupervised Video Object Segmentation

Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, Bo Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8781-8790

The popular unsupervised video object segmentation methods fuse the RGB frame and optical flow via a two-stream network. However, they cannot handle the distracting noises in each input modality, which may vastly deteriorate the model performance. We propose to establish the correspondence between the input modalities while suppressing the distracting signals via optimal structural matching. Given a video frame, we extract the dense local features from the RGB image and optical flow, and treat them as two complex structured representations. The Wasserstein distance is then employed to compute the global optimal flows to transport the features in one modality to the other, where the magnitude of each flow measures the extent of the alignment between two local features. To plug the structural matching into a two-stream network for end-to-end training, we factorize the input cost matrix into small spatial blocks and design a differentiable long-short Sinkhorn module consisting of a long-distant Sinkhorn layer and a short-distant Sinkhorn layer. We integrate the module into a dedicated two-stream network and dub our model TransportNet. Our experiments show that aligning motion-appearance yields the state-of-the-art results on the popular video object segmentation datasets.

****

RDI-Net: Relational Dynamic Inference Networks

Huanyu Wang, Songyuan Li, Shihao Su, Zequn Qin, Xi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4621-4630

Dynamic inference networks, aimed at promoting computational efficiency, go alon

g an adaptive executing path for a given sample. Prevalent methods typically ass ign a router for each convolutional block and sequentially make block-by-block e xecuting decisions, without considering the relations during the dynamic inferen ce. In this paper, we model the relations for dynamic inference from two aspects : the routers and the samples. We design a novel type of router called the relat ional router to model the relations among routers for a given sample. In princip le, the current relational router aggregates the contextual features of precedin g routers by graph convolution and propagates its router features to subsequent ones, making the executing decision for the current block in a long-range manner . Furthermore, we model the relation between samples by introducing a Sample Rel ation Module (SRM), encouraging correlated samples to go along correlated execut ing paths. As a whole, we call our method the Relational Dynamic Inference Netwo rk (RDI-Net). Extensive experiments on CIFAR-10/100 and ImageNet show that RDI-N et achieves state-of-the-art performance and computational cost reduction. Our c ode and models will be made publicly available.

****

Densely Guided Knowledge Distillation Using Multiple Teacher Assistants

Wonchul Son, Jaemin Na, Junyong Choi, Wonjun Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9395-9404

With the success of deep neural networks, knowledge distillation which guides th e learning of a small student network from a large teacher network is being acti vely studied for model compression and transfer learning. However, few studies h ave been performed to resolve the poor learning issue of the student network whe n the student and teacher model sizes significantly differ. In this paper, we pr opose a densely guided knowledge distillation using multiple teacher assistants that gradually decreases the model size to efficiently bridge the large gap betw een the teacher and student networks. To stimulate more efficient learning of th e student network, we guide each teacher assistant to every other smaller teache r assistants iteratively. Specifically, when teaching a smaller teacher assistan t at the next step, the existing larger teacher assistants from the previous ste p are used as well as the teacher network. Moreover, we design stochastic teachi ng where, for each mini-batch, a teacher or teacher assistants are randomly drop ped. This acts as a regularizer to improve the efficiency of teaching of the stu dent network. Thus, the student can always learn salient distilled knowledge fro m the multiple sources. We verified the effectiveness of the proposed method for a classification task using CIFAR-10, CIFAR-100, and ImageNet. We also achieved significant performance improvements with various backbone architectures such a s ResNet, WideResNet, and VGG.

****

Pi-NAS: Improving Neural Architecture Search by Reducing Supernet Training Consi stency Shift

Jiefeng Peng, Jiqi Zhang, Changlin Li, Guangrun Wang, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12354-12364

Recently proposed neural architecture search (NAS) methods co-train billions of architectures in a supernet and estimate their potential accuracy using the netw ork weights detached from the supernet. However, the ranking correlation between the architectures' predicted accuracy and their actual capability is incorrect, which causes the existing NAS methods' dilemma. We attribute this ranking corre lation problem to the supernet training consistency shift, including feature shi ft and parameter shift. Feature shift is identified as dynamic input distributio ns of a hidden layer due to random path sampling. The input distribution dynamic affects the loss descent and finally affects architecture ranking. Parameter sh ift is identified as contradictory parameter updates for a shared layer lay in d ifferent paths in different training steps. The rapidly-changing parameter could not preserve architecture ranking. We address these two shifts simultaneously u sing a nontrivial supernet-\Pi model, called \Pi-NAS. Specifically, we employ a supernet-\Pi model that contains cross-path learning to reduce the feature consi stency shift between different paths. Meanwhile, we adopt a novel nontrivial mea n teacher containing negative samples to overcome parameter shift and model coll

ision. Furthermore, our \Pi-NAS runs in an unsupervised manner, which can search for more transferable architectures. Extensive experiments on ImageNet and a wide range of downstream tasks (e.g., COCO 2017, ADE20K, and Cityscapes) demonstrate the effectiveness and universality of our \Pi-NAS compared to supervised NAS. See Codes: https://github.com/Ernie1/Pi-NAS.
****

ARAPReg: An As-Rigid-As Possible Regularization Loss for Learning Deformable Shape Generators
Qixing Huang, Xiangru Huang, Bo Sun, Zaiwei Zhang, Junfeng Jiang, Chandrajit Bajaj; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5815-5825
This paper introduces an unsupervised loss for training parametric deformation shape generators. The key idea is to enforce the preservation of local rigidity among the generated shapes. Our approach builds on a local approximation of the as-rigid-as possible (or ARAP) deformation energy. We show how to develop the unsupervised loss via a spectral decomposition of the Hessian of the ARAP loss. Our loss nicely decouples pose and shape variations through a robust norm. The loss admits simple closed-form expressions. It is easy to train and can be plugged into any standard generation models, e.g., VAE and GAN. Experimental results show that our approach outperforms existing shape generation approaches considerably across various datasets such as DFAUST, Animal, and Bone.
****

Online Refinement of Low-Level Feature Based Activation Map for Weakly Supervised Object Localization
Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, Linlin Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 132-141
We present a two-stage learning framework for weakly supervised object localization (WSOL). While most previous efforts rely on high-level feature based CAMs (Class Activation Maps), this paper proposes to localize objects using the low-level feature based activation maps. In the first stage, an activation map generator produces activation maps based on the low-level feature maps in the classifier, such that rich contextual object information is included in an online manner. In the second stage, we employ an evaluator to evaluate the activation maps predicted by the activation map generator. Based on this, we further propose a weighted entropy loss, an attentive erasing, and an area loss to drive the activation map generator to substantially reduce the uncertainty of activations between object and background, and explore less discriminative regions. Based on the low-level object information preserved in the first stage, the second stage model gradually generates a well-separated, complete, and compact activation map of object in the image, which can be easily thresholded for accurate localization. Extensive experiments on CUB-200-2011 and ImageNet-1K datasets show that our framework surpasses previous methods by a large margin, which sets a new state-of-the-art for WSOL. Code will be available soon.
****

Grounding Consistency: Distilling Spatial Common Sense for Precise Visual Relationship Detection
Markos Diomataris, Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Maragos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15911-15920
Scene Graph Generators (SGGs) are models that, given an image, build a directed graph where each edge represents a predicted subject predicate object triplet. Most SGGs silently exploit datasets' bias on relationships' context, i.e. its subject and object, to improve recall and neglect spatial and visual evidence, e.g. having seen a glut of data for person wearing shirt, they are overconfident that every person is wearing every shirt. Such imprecise predictions are mainly ascribed to the lack of negative examples for most relationships, fact that obstructs models from meaningfully learning predicates, even those which have ample positive examples. We first present an in-depth investigation of the context bias issue to showcase that all examined state-of-the-art SGGs share the above vulnera

bilities. In response, we propose a semi-supervised scheme that forces predicted triplets to be grounded consistently back to the image, in a closed-loop manner. The developed spatial common sense can be then distilled to a student SGG and substantially enhance its spatial reasoning ability. This Grounding Consistency Distillation (GCD) approach is model-agnostic and profits from the superfluous unlabeled samples to retain the valuable context information and avert memorization of annotations. Furthermore, we ascertain that current metrics disregard unlabeled samples, rendering themselves incapable of reflecting context bias, then we mine and incorporate during evaluation hard-negatives to reformulate precision as a reliable metric. Extensive experimental comparisons exhibit large quantitative - up to 70% relative precision boost on VG200 dataset - and qualitative improvements to prove the significance of our GCD method and our metrics towards refocusing graph generation as a core aspect of scene understanding. Code available at https://github.com/deeplab-ai/grounding-consistent-vrd.

****

Long-Term Temporally Consistent Unpaired Video Translation From Simulated Surgical 3D Data

Dominik Rivoir, Micha Pfeiffer, Reuben Docea, Fiona Kolbinger, Carina Riediger, Jürgen Weitz, Stefanie Speidel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3343-3353

Research in unpaired video translation has mainly focused on short-term temporal consistency by conditioning on neighboring frames. However for transfer from simulated to photorealistic sequences, available information on the underlying geometry offers potential for achieving global consistency across views. We propose a novel approach which combines unpaired image translation with neural rendering to transfer simulated to photorealistic surgical abdominal scenes. By introducing global learnable textures and a lighting-invariant view-consistency loss, our method produces consistent translations of arbitrary views and thus enables long-term consistent video synthesis. We design and test our model to generate video sequences from minimally-invasive surgical abdominal scenes. Because labeled data is often limited in this domain, photorealistic data where ground truth information from the simulated domain is preserved is especially relevant. By extending existing image-based methods to view-consistent videos, we aim to impact the applicability of simulated training and evaluation environments for surgical applications. Code and data: http://opencas.dkfz.de/video-sim2real.

****

Bridging the Gap Between Label- and Reference-Based Synthesis in Multi-Attribute Image-to-Image Translation

Qiusheng Huang, Zhilin Zheng, Xueqi Hu, Li Sun, Qingli Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14628-14637

The image-to-image translation (I2IT) model takes a target label or a reference image as the input, and changes a source into the specified target domain. The two types of synthesis, either label- or reference-based, have substantial differences. Particularly, the label-based synthesis reflects the common characteristics of the target domain, and the reference-based shows the specific style similar to the reference. This paper intends to bridge the gap between them in the task of multi-attribute I2IT. We design the label- and reference-based encoding modules (LEM and REM) to compare the domain differences. They first transfer the source image and target label (or reference) into a common embedding space, by providing the opposite directions through the attribute difference vector. Then the two embeddings are simply fused together to form the latent code S_ rand (or S_ ref), reflecting the domain style differences, which is injected into each layer of the generator by SPADE. To link LEM and REM, so that two types of results benefit each other, we encourage the two latent codes to be close, and set up the cycle consistency between the forward and backward translations on them. Moreover, the interpolation between the S_ rand and S_ ref is also used to synthesize an extra image. Experiments show that label- and reference-based synthesis are indeed mutually promoted, so that we can have the diverse results from LEM, and high quality results with the similar style of the reference.

****

A Broad Study on the Transferability of Visual Representations With Contrastive Learning
Ashraful Islam, Chun-Fu (Richard) Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, Rogerio Feris; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8845-8855
Tremendous progress has been made in visual representation learning, notably with the recent success of self-supervised contrastive learning methods. Supervised contrastive learning has also been shown to outperform its cross-entropy counterparts by leveraging labels for choosing where to contrast. However, there has been little work to explore the transfer capability of contrastive learning to a different domain. In this paper, we conduct a comprehensive study on the transferability of learned representations of different contrastive approaches for linear evaluation, full-network transfer, and few-shot recognition on 12 downstream datasets from different domains, and object detection tasks on MSCOCO and VOC0712. The results show that the contrastive approaches learn representations that are easily transferable to a different downstream task. We further observe that the joint objective of self-supervised contrastive loss with cross-entropy/supervised-contrastive loss leads to better transferability of these models over their supervised counterparts. Our analysis reveals that the representations learned from the contrastive approaches contain more low/mid-level semantics than cross-entropy models, which enables them to quickly adapt to a new task. Our codes and models will be publicly available to facilitate future research on transferability of visual representations.
****
TempNet: Online Semantic Segmentation on Large-Scale Point Cloud Series
Yunsong Zhou, Hongzi Zhu, Chunqin Li, Tiankai Cui, Shan Chang, Minyi Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7118-7127
Online semantic segmentation on a time series of point cloud frames is an essential task in autonomous driving. Existing models focus on single-frame segmentation, which cannot achieve satisfactory segmentation accuracy and offer unstably flicker among frames. In this paper, we propose a light-weight semantic segmentation framework for large-scale point cloud series, called TempNet, which can improve both the accuracy and the stability of existing semantic segmentation models by combining a novel frame aggregation scheme. To be computational cost efficient, feature extraction and aggregation are only conducted on a small portion of key frames via a temporal feature aggregation (TFA) network using an attentional pooling mechanism, and such enhanced features are propagated to the intermediate non-key frames. To avoid information loss from non-key frames, a partial feature update (PFU) network is designed to partially update the propagated features with the local features extracted on a non-key frame if a large disparity between the two is quickly assessed. As a result, consistent and information-rich features can be obtained for each frame. We implement TempNet on five state-of-the-art (SOTA) point cloud segmentation models and conduct extensive experiments on the SemanticKITTI dataset. Results demonstrate that TempNet outperforms SOTA competitors by wide margins with little extra computational cost.
****
Bayesian Deep Basis Fitting for Depth Completion With Uncertainty
Chao Qu, Wenxin Liu, Camillo J. Taylor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16147-16157
In this work we investigate the problem of uncertainty estimation for image-guided depth completion. We extend Deep Basis Fitting (DBF) for depth completion within a Bayesian evidence framework to provide calibrated per-pixel variance. The DBF approach frames the depth completion problem in terms of a network that produces a set of low-dimensional depth bases and a differentiable least-squares fitting module that computes the basis weights using the sparse depths. By adopting a Bayesian treatment, our Bayesian Deep Basis Fitting (BDBF) approach is able to 1) predict high-quality uncertainty estimates and 2) enable depth completion with few or no sparse measurements. We conduct controlled experiments to compare BDBF against commonly used techniques for uncertainty estimation under various s

cenarios. Results show that our method produces better uncertainty estimates wit
h accurate depth prediction.
****

Query Adaptive Few-Shot Object Detection With Heterogeneous Graph Convolutional
Networks
Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, Shih-Fu Chang; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 32
63-3272
Few-shot object detection (FSOD) aims to detect never-seen objects using few exa
mples. This field sees recent improvement owing to the meta-learning techniques
by learning how to match between the query image and few-shot class examples, su
ch that the learned model can generalize to few-shot novel classes. However, cur
rently, most of the meta-learning-based methods perform parwise matching between
 query image regions (usually proposals) and novel classes separately, therefore
 failing to take into account multiple relationships among them. In this paper,
we propose a novel FSOD model using heterogeneous graph convolutional networks.
Through efficient message passing among all the proposal and class nodes with th
ree different types of edges, we could obtain context-aware proposal features an
d query-adaptive, multiclass-enhanced prototype representations for each class,
which could help promote the pairwise matching and improve final FSOD accuracy.
Extensive experimental results show that our proposed model, denoted as QA-FewDe
t, outperforms the current state-of-the-art approaches on the PASCAL VOC and MSC
OCO FSOD benchmarks under different shots and evaluation metrics.
****

ResRep: Lossless CNN Pruning via Decoupling Remembering and Forgetting
Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, Guig
uang Ding; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2021, pp. 4510-4520
We propose ResRep, a novel method for lossless channel pruning (a.k.a. filter pr
uning), which slims down a CNN by reducing the width (number of output channels)
 of convolutional layers. Inspired by the neurobiology research about the indepe
ndence of remembering and forgetting, we propose to re-parameterize a CNN into t
he remembering parts and forgetting parts, where the former learn to maintain th
e performance and the latter learn to prune. Via training with regular SGD on th
e former but a novel update rule with penalty gradients on the latter, we realiz
e structured sparsity. Then we equivalently merge the remembering and forgetting
 parts into the original architecture with narrower layers. In this sense, ResRe
p can be viewed as a successful application of Structural Re-parameterization. S
uch a methodology distinguishes ResRep from the traditional learning-based pruni
ng paradigm that applies a penalty on parameters to produce sparsity, which may
suppress the parameters essential for the remembering. ResRep slims down a stand
ard ResNet-50 with 76.15% accuracy on ImageNet to a narrower one with only 45% F
LOPs and no accuracy drop, which is the first to achieve lossless pruning with s
uch a high compression ratio. The code and models are at https://github.com/Ding
XiaoH/ResRep.
****

P2-Net: Joint Description and Detection of Local Features for Pixel and Point Ma
tching
Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu,
Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, Andrew Markham; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16004-1
6013
Accurately describing and detecting 2D and 3D keypoints is crucial to establishi
ng correspondences across images and point clouds. Despite a plethora of learnin
g-based 2D or 3D local feature descriptors and detectors having been proposed, t
he derivation of a shared descriptor and joint keypoint detector that directly m
atches pixels and points remains under-explored by the community. This work take
s the initiative to establish fine-grained correspondences between 2D images and
 3D point clouds. In order to directly match pixels and points, a dual fully con
volutional framework is presented that maps 2D and 3D inputs into a shared laten

t representation space to simultaneously describe and detect keypoints. Furtherm ore, an ultra-wide reception mechanism and a novel loss function are designed to mitigate the intrinsic information variations between pixel and point local reg ions. Extensive experimental results demonstrate that our framework shows compet itive performance in fine-grained matching between images and point clouds and a chieves state-of-the-art results for the task of indoor visual localization. Our source code will be available at [no-name-for-blind-review].
****

Generalize Then Adapt: Source-Free Domain Adaptive Semantic Segmentation
Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, R. Venkatesh Ba bu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 7046-7056
Unsupervised domain adaptation (DA) has gained substantial interest in semantic segmentation. However, almost all prior arts assume concurrent access to both la beled source and unlabeled target, making them unsuitable for scenarios demandin g source-free adaptation. In this work, we enable source-free DA by partitioning the task into two: a) source-only domain generalization and b) source-free targ et adaptation. Towards the former, we provide theoretical insights to develop a multi-head framework trained with a virtually extended multi-source dataset, aim ing to balance generalization and specificity. Towards the latter, we utilize th e multi-head framework to extract reliable target pseudo-labels for self-trainin g. Additionally, we introduce a novel conditional prior-enforcing auto-encoder t hat discourages spatial irregularities, thereby enhancing the pseudo-label quali ty. Experiments on the standard GTA5-to-Cityscapes and SYNTHIA-to-Cityscapes ben chmarks show our superiority even against the non-source-free prior-arts. Furthe r, we show our compatibility with online adaptation enabling deployment in a seq uentially changing environment.
****

Cross-Modality Person Re-Identification via Modality Confusion and Center Aggreg ation
Xin Hao, Sanyuan Zhao, Mang Ye, Jianbing Shen; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 16403-16412
Cross-modality person re-identification is a challenging task due to large cross -modality discrepancy and intra-modality variations. Currently, most existing me thods focus on learning modality-specific or modality-shareable features by usin g the identity supervision or modality label. Different from existing methods, t his paper presents a novel Modality Confusion Learning Network (MCLNet). Its bas ic idea is to confuse two modalities, ensuring that the optimization is explicit ly concentrated on the modality-irrelevant perspective. Specifically, MCLNet is designed to learn modality-invariant features by simultaneously minimizing inter -modality discrepancy while maximizing cross-modality similarity among instances in a single framework. Furthermore, an identity-aware marginal center aggregati on strategy is introduced to extract the centralization features, while keeping diversity with a marginal constraint. Finally, we design a camera-aware learning scheme to enrich the discriminability. Extensive experiments on SYSU-MM01 and R egDB datasets show that MCLNet outperforms the state-of-the-art by a large margi n. On the large-scale SYSU-MM01 dataset, our model can achieve 65.40% and 61.98% in terms of Rank-1 accuracy and mAP value.
****

T-Net: Effective Permutation-Equivariant Network for Two-View Correspondence Lea rning
Zhen Zhong, Guobao Xiao, Linxin Zheng, Yan Lu, Jiayi Ma; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 1950-1959
We develop a conceptually simple, flexible, and effective framework (named T-Net) for two-view correspondence learning. Given a set of putative correspondences, we reject outliers and regress the relative pose encoded by the essential matri x, by an end-to-end framework, which is consisted of two novel structures: "-" s tructure and "|" structure. "-" structure adopts an iterative strategy to learn correspondence features. "|" structure integrates all the features of the iterat ions and outputs the correspondence weight. In addition, we introduce Permutatio

n-Equivariant Context Squeeze-and-Excitation module, an adapted version of SE mo dule, to process sparse correspondences in a permutation-equivariant way and cap ture both global and channel-wise contextual information. Extensive experiments on outdoor and indoor scenes show that the proposed T-Net achieves state-of-the- art performance. On outdoor scenes (YFCC100M dataset), T-Net achieves an mAP of 52.28%, a 34.22% precision increase from the best-published result (38.95%). On indoor scenes (SUN3D dataset), T-Net (19.71%) obtains a 21.82% precision increas e from the best-published result (16.18%).
****

Temporal Cue Guided Video Highlight Detection With Low-Rank Audio-Visual Fusion
Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, Guang Yang; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7950-7959
Video highlight detection plays an increasingly important role in social media c ontent filtering, however, it remains highly challenging to develop automated vi deo highlight detection methods because of the lack of temporal annotations (i.e ., where the highlight moments are in long videos) for supervised learning. In t his paper, we propose a novel weakly supervised method that can learn to detect highlights by mining video characteristics with video level annotations (topic t ags) only. Particularly, we exploit audio-visual features to enhance video repre sentation and take temporal cues into account for improving detection performanc e. Our contributions are threefold: 1) we propose an audio-visual tensor fusion mechanism that efficiently models the complex association between two modalities while reducing the gap of the heterogeneity between the two modalities; 2) we i ntroduce a novel hierarchical temporal context encoder to embed local temporal c lues in between neighboring segments; 3) finally, we alleviate the gradient vani shing problem theoretically during model optimization with attention-gated insta nce aggregation. Extensive experiments on two benchmark datasets (YouTube Highli ghts and TVSum) have demonstrated our method outperforms other state-of-the-art methods with remarkable improvements.
****

Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models
Linjie Li, Jie Lei, Zhe Gan, Jingjing Liu; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 2042-2051
Benefiting from large-scale pre-training, we have witnessed significant performa nce boost on the popular Visual Question Answering (VQA) task. Despite rapid pro gress, it remains unclear whether these state-of-the-art (SOTA) models are robus t when encountering examples in the wild. To study this, we introduce Adversaria l VQA, a new large-scale VQA benchmark, collected iteratively via an adversarial human-and-model-in-the-loop procedure. Through this new benchmark, we discover several interesting findings. (i) Surprisingly, we find that during dataset coll ection, non-expert annotators can easily attack SOTA VQA models successfully. (i i) Both large-scale pre-trained models and adversarial training methods achieve far worse performance on the new benchmark than over standard VQA v2 dataset, re vealing the fragility of these models while demonstrating the effectiveness of o ur adversarial dataset. (iii) When used for data augmentation, our dataset can e ffectively boost model performance on other robust VQA benchmarks. We hope our A dversarial VQA dataset can shed new light on robustness study in the community a nd serve as a valuable benchmark for future work.
****

S3VAADA: Submodular Subset Selection for Virtual Adversarial Active Domain Adapt ation
Harsh Rangwani, Arihant Jain, Sumukh K Aithal, R. Venkatesh Babu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7516 -7525
Unsupervised domain adaptation (DA) methods have focused on achieving maximal pe rformance through aligning features from source and target domains without using labeled data in the target domain. Whereas, in the real-world scenario's it mig ht be feasible to get labels for a small proportion of target data. In these sce narios, it is important to select maximally-informative samples to label and fin

d an effective way to combine them with the existing knowledge from source data. Towards achieving this, we propose S^3VAADA which i) introduces a novel submodular criterion to select a maximally informative subset to label and ii) enhances a cluster-based DA procedure through novel improvements to effectively utilize all the available data for improving generalization on target. Our approach consistently outperforms the competing state-of-the-art approaches on datasets with varying degrees of domain shifts. The project page with additional details is available here: https://sites.google.com/iisc.ac.in/s3vaada-iccv2021.
****

Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation
Jiabo Huang, Yang Liu, Shaogang Gong, Hailin Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7199-7208
Video activity localisation has recently attained increasing attention due to its practical values in automatically localising the most salient visual segments corresponding to their language descriptions (sentences) from untrimmed and unstructured videos. For supervised model training, a temporal annotation of both the start and end time index of each video segment for a sentence (a video moment) must be given. This is not only very expensive but also sensitive to ambiguity and subjective annotation bias, a much harder task than image labelling. In this work, we develop a more accurate weakly-supervised solution by introducing Cross-Sentence Relations Mining (CRM) in video moment proposal generation and matching when only a paragraph description of activities without per-sentence temporal annotation is available. Specifically, we explore two cross-sentence relational constraints: (1) Temporal ordering and (2) semantic consistency among sentences in a paragraph description of video activities. Existing weakly-supervised techniques only consider within-sentence video segment correlations in training without considering cross-sentence paragraph context. This can mislead due to ambiguous expressions of individual sentences with visually indiscriminate video moment proposals in isolation. Experiments on two publicly available activity localisation datasets show the advantages of our approach over the state-of-the-art weakly supervised methods, especially so when the video activity descriptions become more complex.
****

StructDepth: Leveraging the Structural Regularities for Self-Supervised Indoor Depth Estimation
Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, Wenxian Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12663-12673
Self-supervised monocular depth estimation has achieved impressive performance on outdoor datasets. Its performance however degrades notably in indoor environments because of the lack of textures. Without rich textures, the photometric consistency is too weak to train a good depth network. Inspired by the early works on indoor modeling, we leverage the structural regularities exhibited in indoor scenes, to train a better depth network. Specifically, we adopt two extra supervisory signals for self-supervised training: 1) the Manhattan normal constraint and 2) the co-planar constraint. The Manhattan normal constraint enforces the major surfaces (the floor, ceiling, and walls) to be aligned with dominant directions. The co-planar constraint states that the 3D points be well fitted by a plane if they are located within the same planar region. To generate the supervisory signals, we adopt two components to classify the major surface normal into dominant directions and detect the planar regions on the fly during training. As the predicted depth becomes more accurate after more training epochs, the supervisory signals also improve and in turn feedback to obtain a better depth model. Through extensive experiments on indoor benchmark datasets, the results show that our network outperforms the state-of-the-art methods. The source code is available at https://github.com/SJTU-ViSYS/StructDepth.
****

Feature Interactive Representation for Point Cloud Registration
Bingli Wu, Jie Ma, Gaojie Chen, Pei An; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5530-5539
Point cloud registration is the process of using the common structures in two po

int clouds to splice them together. To find out these common structures and make these structures match more accurately, we investigate the direction of interacting information of the source and target point clouds. To this end, we propose a Feature Interactive Representation learning Network (FIRE-Net), which can explore feature interaction among the source and target point clouds from different levels. Specifically, we first introduce a Combined Feature Encoder (CFE) based on feature interaction intra point cloud. CFE extracts interactive features intra each point cloud and combines them to enhance the ability of the network to describe the local geometric structure. Then, we propose a feature interaction mechanism inter point clouds which includes a Local Interaction Unit (LIU) and a Global Interaction Unit (GIU). The former is used to interact information between point pairs across two point clouds, thus the point features in one point cloud and its similar point features in another point cloud can be aware of each other. The latter is applied to change the per-point features depending on the global cross information of two point clouds, thus one point cloud has the global perception of another. Extensive experiments on partially overlapping point cloud registration show that our method achieves state-of-the-art performance.
****

Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis
Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, Weidong Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 915-924
Discrete point cloud objects lack sufficient shape descriptors of 3D geometries. In this paper, we present a novel method for aggregating hypothetical curves in point clouds. Sequences of connected points (curves) are initially grouped by taking guided walks in the point clouds, and then subsequently aggregated back to augment their point-wise features. We provide an effective implementation of the proposed aggregation strategy including a novel curve grouping operator followed by a curve aggregation operator. Our method was benchmarked on several point cloud analysis tasks where we achieved the state-of-the-art classification accuracy of 94.2% on the ModelNet40 classification task, instance IoU of 86.8% on the ShapeNetPart segmentation task and cosine error of 0.11 on the ModelNet40 normal estimation task.
****

LSG-CPD: Coherent Point Drift With Local Surface Geometry for Point Cloud Registration
Weixiao Liu, Hongtao Wu, Gregory S. Chirikjian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15293-15302
Probabilistic point cloud registration methods are becoming more popular because of their robustness. However, unlike point-to-plane variants of iterative closest point (ICP) which incorporate local surface geometric information such as surface normals, most probabilistic methods (e.g., coherent point drift (CPD)) ignore such information and build Gaussian mixture models (GMMs) with isotropic Gaussian covariances. This results in sphere-like GMM components which only penalize the point-to-point distance between the two point clouds. In this paper, we propose a novel method called CPD with Local Surface Geometry (LSG-CPD) for rigid point cloud registration. Our method adaptively adds different levels of point-to-plane penalization on top of the point-to-point penalization based on the flatness of the local surface. This results in GMM components with anisotropic covariances. We formulate point cloud registration as a maximum likelihood estimation (MLE) problem and solve it with the Expectation-Maximization (EM) algorithm. In the E step, we demonstrate that the computation can be recast into simple matrix manipulations and efficiently computed on a GPU. In the M step, we perform an unconstrained optimization on a matrix Lie group to efficiently update the rigid transformation of the registration. The proposed method outperforms state-of-the-art algorithms in terms of accuracy and robustness on various datasets captured with range scanners, RGBD cameras, and LiDARs. Also, it is significantly faster than modern implementations of CPD. The source code is available at https://github.com/ChirikjianLab/LSG-CPD.git.
****

ISD: Self-Supervised Learning by Iterative Similarity Distillation
Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, Hamed Pirsiavash; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9609-9618
Recently, contrastive learning has achieved great results in self-supervised learning, where the main idea is to pull two augmentations of an image (positive pairs) closer compared to other random images (negative pairs). We argue that not all negative images are equally negative. Hence, we introduce a self-supervised learning algorithm where we use a soft similarity for the negative images rather than a binary distinction between positive and negative pairs. We iteratively distill a slowly evolving teacher model to the student model by capturing the similarity of a query image to some random images and transferring that knowledge to the student. Specifically, our method should handle unbalanced and unlabeled data better than existing contrastive learning methods, because the randomly chosen negative set might include many samples that are semantically similar to the query image. In this case, our method labels them as highly similar while standard contrastive methods label them as negatives. Our method achieves comparable results to the state-of-the-art models. Our code is available here: https://github.com/UMBCvision/ISD
****

An Empirical Study of the Collapsing Problem in Semi-Supervised 2D Human Pose Estimation
Rongchang Xie, Chunyu Wang, Wenjun Zeng, Yizhou Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11240-11249
The state-of-the-art semi-supervised learning models are consistency-based which learn about unlabeled images by maximizing the similarity between different augmentations of an image. But when we apply the methods to human pose estimation which has extremely imbalanced class distribution, the models often collapse and predict every pixel in unlabeled images as background. This is because the decision boundary may pass through the high-density area of the minor class so more and more pixels are gradually mis-classified as the background class. In this work, we present a surprisingly simple approach to drive the model to learn in the correct direction. For each image, it composes a pair of easy and hard augmentations and uses the more accurate predictions on the easy image to teach the network to learn about the hard one. The accuracy superiority of teaching signals allows the network to be "monotonically" improved which effectively avoids collapsing. We apply our method to recent pose estimators and find that they achieve significantly better performances than their supervised counterparts on three public datasets.
****

Self-Supervised Neural Networks for Spectral Snapshot Compressive Imaging
Ziyi Meng, Zhenming Yu, Kun Xu, Xin Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2622-2631
We consider using untrained neural networks to solve the reconstruction problem of snapshot compressive imaging (SCI), which uses a two-dimensional (2D) detector to capture a high-dimensional (usually 3D) data-cube in a compressed manner. Various SCI systems have been built in recent years to capture data such as high-speed videos, hyperspectral images, and the state-of-the-art reconstruction is obtained by the deep neural networks. However, most of these networks are trained in an end-to-end manner by a large amount of corpus with sometimes simulated ground truth, measurement pairs. In this paper, inspired by the untrained neural networks such as deep image priors (DIP) and deep decoders, we develop a framework by integrating DIP into the plug-and-play regime, leading to a self-supervised network for spectral SCI reconstruction. Extensive synthetic and real data results show that the proposed algorithm without training is capable of achieving competitive results to the training based networks. Furthermore, by integrating the proposed method with a pre-trained deep denoising prior, we have achieved higher performance than existing state-of-the-art.
****

Group-Aware Contrastive Regression for Action Quality Assessment

Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7919-7928

Assessing action quality is challenging due to the subtle differences between videos and large variations in scores. Most existing approaches tackle this problem by regressing a quality score from a single video, suffering a lot from the large inter-video score variations. In this paper, we show that the relations among videos can provide important clues for more accurate action quality assessment during both training and inference. Specifically, we reformulate the problem of action quality assessment as regressing the relative scores with reference to another video that has shared attributes (e.g. category and difficulty), instead of learning unreferenced scores. Following this formulation, we propose a new contrastive regression (CoRe) framework to learn the relative scores by pair-wise comparison, which highlights the differences between videos and guides the models to learn the key hints for assessment. In order to further exploit the relative information between two videos, we devise a group-aware regression tree to convert the conventional score regression into two easier sub-problems: coarse-to-fine classification and regression in small intervals. To demonstrate the effectiveness of CoRe, we conduct extensive experiments on three mainstream AQA datasets including AQA-7, MTL-AQA, and JIGSAWS. Our approaches outperform previous methods by a large margin and establish new state-of-the-art on all three benchmarks.

****

The Road To Know-Where: An Object-and-Room Informed Sequential BERT for Indoor Vision-Language Navigation

Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1655-1664

Vision-and-Language Navigation (VLN) requires an agent to find a path to a remote location on the basis of natural-language instructions and a set of photo-realistic panoramas. Most existing methods take the words in the instructions and the discrete views of each panorama as the minimal unit of encoding. However, this requires a model to match different nouns (e.g., TV, table) against the same input view feature. In this work, we propose an object-informed sequential BERT to encode visual perceptions and linguistic instructions at the same fine-grained level, namely objects and words. Our sequential BERT also enables the visual-textual clues to be interpreted in light of the temporal context, which is crucial to multi-round VLN tasks. Additionally, we enable the model to identify the relative direction (e.g., left/right/front/back) of each navigable location and the room type (e.g., bedroom, kitchen) of its current and final navigation goal, as such information is widely mentioned in instructions implying the desired next and final locations. We thus enable the model to know-where the objects lie in the images, and to know-where they stand in the scene. Extensive experiments demonstrate the effectiveness compared against several state-of-the-art methods on three indoor VLN tasks: REVERIE, NDH, and R2R. Project repository: https://github.com/YuankaiQi/ORIST

****

Support-Set Based Cross-Supervision for Video Grounding

Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, Xinbo Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11573-11582

Current approaches for video grounding propose kinds of complex architectures to capture the video-text relations, and have achieved impressive improvements. However, it is hard to learn the complicated multi-modal relations by only architecture designing in fact. In this paper, we introduce a novel Support-set Based Cross-Supervision (Sscs) module which can improve existing methods during training phase without extra inference cost. The contrastive objective aims to learn effective representations by contrastive learning, while the caption objective can train a powerful video encoder supervised by texts. Due to the co-existence of some visual entities in both ground-truth and background intervals, i.e., mutual exclusion, naively contrastive learning is unsuitable to video grounding. We ad

dress the problem by boosting the cross-supervision with the support-set concept , which collects visual information from the whole video and eliminates the mutu al exclusion of entities. Combined with the original objective, Sscs can enhance the abilities of multi-modal relation modeling for existing approaches. We exte nsively evaluate Sscs on three challenging datasets, and show that our method ca n improves current state-of-the-art methods by large margins, especially 6.35% i n terms of R1@0.5 on Charades-STA.
****

Sampling Network Guided Cross-Entropy Method for Unsupervised Point Cloud Regist ration
Haobo Jiang, Yaqi Shen, Jin Xie, Jun Li, Jianjun Qian, Jian Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6128 -6137
In this paper, by modeling the point cloud registration task as a Markov decisio n process, we propose an end-to-end deep model embedded with the cross-entropy m ethod (CEM) for unsupervised 3D registration. Our model consists of a sampling n etwork module and a differentiable CEM module. In our sampling network module, g iven a pair of point clouds, the sampling network learns a prior sampling distri bution over the transformation space. The learned sampling distribution can be u sed as a "good"" initialization of the differentiable CEM module. In our differe ntiable CEM module, we first propose a maximum consensus criterion based alignme nt metric as the reward function for the point cloud registration task. Based on the reward function, for each state, we then construct a fused score function t o evaluate the sampled transformations, where we weight the current and future r ewards of the transformations. Particularly, the future rewards of the sampled t ransforms are obtained by performing the iterative closest point (ICP) algorithm on the transformed state. By selecting the top-k transformations with the highe st scores, we iteratively update the sampling distribution. Furthermore, in orde r to make the CEM differentiable, we use the sparsemax function to replace the h ard top-k selection. Finally, we formulate a Geman-McClure estimator based loss to train our end-to-end registration model. Extensive experimental results demon strate the good registration performance of our method on benchmark datasets.
****

Voxel-Based Network for Shape Completion by Leveraging Edge Generation
Xiaogang Wang, Marcelo H Ang, Gim Hee Lee; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 13189-13198
Deep learning technique has yielded significant improvements in point cloud comp letion with the aim of completing missing object shapes from partial inputs. How ever, most existing methods fail to recover realistic structures due to over-smo othing of fine-grained details. In this paper, we develop a voxel-based network for point cloud completion by leveraging edge generation (VE-PCN). We first embe d point clouds into regular voxel grids, and then generate complete objects with the help of the hallucinated shape edges. This decoupled architecture together with a multi-scale grid feature learning is able to generate more realistic on-s urface details. We evaluate our model on the publicly available completion datas ets and show that it outperforms existing state-of-the-art approaches quantitati vely and qualitatively. Our source code is available at https://github.com/xiaog angw/VE-PCN.
****

THUNDR: Transformer-Based 3D Human Reconstruction With Markers
Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul S ukthankar, Cristian Sminchisescu; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2021, pp. 12971-12980
We present THUNDR, a transformer-based deep neural network methodology to recons truct the 3d pose and shape of people, given monocular RGB images. Key to our me thodology is an intermediate 3d marker representation, where we aim to combine t he predictive power of model-free-output architectures and the regularizing, ant hropometrically-preserving properties of a statistical human surface model like GHUM---a recently introduced, expressive full body statistical 3d human model, t rained end-to-end. Our novel transformer-based prediction pipeline can focus on

image regions relevant to the task, supports self-supervised regimes, and ensures that solutions are consistent with human anthropometry. We show state-of-the-art results on Human3.6M and 3DPW, for both the fully-supervised and the self-supervised models, for the task of inferring 3d human shape, joint positions, and global translation. Moreover, we observe very solid 3d reconstruction performance for difficult human poses collected in the wild.
****

OadTR: Online Action Detection With Transformers
Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7565-7575
Most recent approaches for online action detection tend to apply Recurrent Neural Network (RNN) to capture long-range temporal structure. However, RNN suffers from non-parallelism and gradient vanishing, hence it is hard to be optimized. In this paper, we propose a new encoder-decoder framework based on Transformers, named OadTR, to tackle these problems. The encoder attached with a task token aims to capture the relationships and global interactions between historical observations. The decoder extracts auxiliary information by aggregating anticipated future clip representations. Therefore, OadTR can recognize current actions by encoding historical information and predicting future context simultaneously. We extensively evaluate the proposed OadTR on three challenging datasets: HDD, TVSeries, and THUMOS14. The experimental results show that OadTR achieves higher training and inference speeds than current RNN based approaches, and significantly outperforms the state-of-the-art methods in terms of both mAP and mcAP. Code is available at https: //github.com/wangxiang1230/OadTR.
****

Instance-Level Image Retrieval Using Reranking Transformers
Fuwen Tan, Jiangbo Yuan, Vicente Ordonez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12105-12115
Instance-level image retrieval is the task of searching in a large database for images that match an object in a query image. To address this task, systems usually rely on a retrieval step that uses global image descriptors, and a subsequent step that performs domain-specific refinements or reranking by leveraging operations such as geometric verification based on local features. In this work, we propose Reranking Transformers (RRTs) as a general model to incorporate both local and global features to rerank the matching images in a supervised fashion and thus replace the relatively expensive process of geometric verification. RRTs are lightweight and can be easily parallelized so that reranking a set of top matching results can be performed in a single forward-pass. We perform extensive experiments on the Revisited Oxford and Paris datasets, and the Google Landmarks v2 dataset, showing that RRTs outperform previous reranking approaches while using much fewer local descriptors. Moreover, we demonstrate that, unlike existing approaches, RRTs can be optimized jointly with the feature extractor, which can lead to feature representations tailored to downstream tasks and further accuracy improvements. The code and trained models are publicly available at https://github.com/uvavision/RerankingTransformer.
****

Mutual-Complementing Framework for Nuclei Detection and Segmentation in Pathology Image
Zunlei Feng, Zhonghua Wang, Xinchao Wang, Yining Mao, Thomas Li, Jie Lei, Yuexuan Wang, Mingli Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4036-4045
Detection and segmentation of nuclei are fundamental analysis operations in pathology images, the assessments derived from which serve as the gold standard for cancer diagnosis. Manual segmenting nuclei is expensive and time-consuming. What's more, accurate segmentation detection of nuclei can be challenging due to the large appearance variation, conjoined and overlapping nuclei, and serious degeneration of histological structures. Supervised methods highly rely on massive annotated samples. The existing two unsupervised methods are prone to failure on degenerated samples. This paper proposes a Mutual-Complementing Framework (MCF) f

or nuclei detection and segmentation in pathology images. Two branches of MCF ar e trained in the mutual-complementing manner, where the detection branch complem ents the pseudo mask of the segmentation branch, while the progressive trained s egmentation branch complements the missing nucleus templates through calculating the mask residual between the predicted mask and detected result. In the detect ion branch, two response map fusion strategies and gradient direction based post processing are devised to obtain the optimal detection response. Furthermore, th e confidence loss combined with the synthetic samples and self-finetuning is ado pted to train the segmentation network with only high confidence areas. Extensiv e experiments demonstrate that MCF achieves comparable performance with only a f ew nucleus patches as supervision. Especially, MCF possesses good robustness (on ly dropping by about 6%) on degenerated samples, which are critical and common c ases in clinical diagnosis.

****

Accelerating Atmospheric Turbulence Simulation via Learned Phase-to-Space Transf orm

Zhiyuan Mao, Nicholas Chimitt, Stanley H. Chan; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2021, pp. 14759-14768

Fast and accurate simulation of imaging through atmospheric turbulence is essent ial for developing turbulence mitigation algorithms. Recognizing the limitations of previous approaches, we introduce a new concept known as the phase-to-space (P2S) transform to significantly speed up the simulation. P2S is built upon thre e ideas: (1) reformulating the spatially varying convolution as a set of invaria nt convolutions with basis functions, (2) learning the basis function via the kn own turbulence statistics models, (3) implementing the P2S transform via a light -weight network that directly converts the phase representation to spatial repre sentation. The new simulator offers 300x - 1000x speed up compared to the mainst ream split-step simulators while preserving the essential turbulence statistics.

****

Graph Constrained Data Representation Learning for Human Motion Segmentation

Mariella Dimiccoli, Lluís Garrido, Guillem Rodriguez-Corominas, Herwig Wendt; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 1460-1469

Recently, transfer subspace learning based approaches have shown to be a valid a lternative to unsupervised subspace clustering and temporal data clustering for human motion segmentation (HMS). These approaches leverage prior knowledge from a source domain to improve clustering performance on a target domain, and curren tly they represent the state of the art in HMS. Bucking this trend, in this pape r, we propose a novel unsupervised model that learns a representation of the dat a and digs clustering information from the data itself. Our model is reminiscent of temporal subspace clustering, but presents two critical differences. First, we learn an auxiliary data matrix that can deviate from the initial data, hence confers more degrees of freedom to the coding matrix. Second, we introduce a reg ularization term for this auxiliary data matrix that preserves the local geometr ical structure present in the high-dimensional space. The proposed model is effi ciently optimized by using an original Alternating Direction Method of Multiplie rs (ADMM) formulation allowing to learn jointly the auxiliary data representatio n, a nonnegative dictionary and a coding matrix. Experimental results on four be nchmark datasets for HMS demonstrate that our approach achieves significantly be tter clustering performance then state-of-the-art methods, including both unsupe rvised and more recent semi-supervised transfer learning approaches.

****

Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recogn ition

Jian Jia, Xiaotang Chen, Kaiqi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 962-971

While recent studies on pedestrian attribute recognition have shown remarkable p rogress in leveraging complicated networks and attention mechanisms, most of the m neglect the inter-image relations and an important prior: spatial consistency and semantic consistency of attributes under surveillance scenarios. The spatial

locations of the same attribute should be consistent between different pedestrian images, e.g., the "hat" attribute and the "boots" attribute are always located at the top and bottom of the picture respectively. In addition, the inherent semantic feature of the "hat" attribute should be consistent, whether it is a baseball cap, beret, or helmet. To fully exploit inter-image relations and aggregate human prior in the model learning process, we construct a Spatial and Semantic Consistency (SSC) framework that consists of two complementary regularizations to achieve spatial and semantic consistency for each attribute. Specifically, we first propose a spatial consistency regularization to focus on reliable and stable attribute-related regions. Based on the precise attribute locations, we further propose a semantic consistency regularization to extract intrinsic and discriminative semantic features. We conduct extensive experiments on popular benchmarks including PA100K, RAP, and PETA. Results show that the proposed method performs favorably against state-of-the-art methods without increasing parameters.
****

Learning To Stylize Novel Views
Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13869-13878
We tackle a 3D scene stylization problem -- generating stylized images of a scene from arbitrary novel views given a set of images of the same scene and a reference image of the desired style as inputs. Direct solution of combining novel view synthesis and stylization approaches lead to results that are blurry or not consistent across different views. We propose a point cloud-based method for consistent 3D scene stylization. First, we construct the point cloud by back-projecting the image features to the 3D space. Second, we develop point cloud aggregation modules to gather the style information of the 3D scene, and then modulate the features in the point cloud with a linear transformation matrix. Finally, we project the transformed features to 2D space to obtain the novel views. Experimental results on two diverse datasets of real-world scenes validate that our method generates consistent stylized novel view synthesis results against other alternative approaches.
****

Morphable Detector for Object Detection on Demand
Xiangyun Zhao, Xu Zou, Ying Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4771-4780
Many emerging applications of intelligent robots need to explore and understand new environments, where it is desirable to detect objects of novel categories on the fly with minimum online efforts. This is an object detection on demand (ODOD) task. It is challenging, because it is impossible to annotate large data on the fly, and the embedded systems are usually unable to perform back-propagation which is essential for training. Most existing few-shot detection methods are confronted here as they need extra training. We propose a novel morphable detector (MD), that simply "morphs"" some of its changeable parameters online estimated from the few samples, so as to detect novel categories without any extra training. The MD has two sets of parameters, one for the feature embedding and the other for category representation(called "prototypes""). Each category is associated with a hidden prototype to be learned by integrating the visual and semantic embeddings. The learning of the MD is based on the alternate learning of the feature embedding and the prototypes in an EM-like approach which allows the recovery of an unknown prototype from a few samples of a novel category. Once an MD is learned, it is able to use a few samples of a novel category to directly compute its prototype to fulfill the online morphing process. We have shown the superiority of the MD in Pascal, COCO and FSOD datasets.
****

Stacked Homography Transformations for Multi-View Pedestrian Detection
Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6049-6057
Multi-view pedestrian detection aims to predict a bird's eye view (BEV) occupanc

y map from multiple camera views. This task is confronted with two challenges: h ow to establish the 3D correspondences from views to the BEV map and how to asse mble occupancy information across views. In this paper, we propose a novel Stack ed HOmography Transformations (SHOT) approach, which is motivated by approximati ng projections in 3D world coordinates via a stack of homographies. We first con struct a stack of transformations for projecting views to the ground plane at di fferent height levels. Then we design a soft selection module so that the networ k learns to predict the likelihood of the stack of transformations. Moreover, we provide an in-depth theoretical analysis on constructing SHOT and how well SHOT approximates projections in 3D world coordinates. SHOT is empirically verified to be capable of estimating accurate correspondences from individual views to th e BEV map, leading to new state-of-the-art performance on standard evaluation be nchmarks.

****

Env-QA: A Video Question Answering Benchmark for Comprehensive Understanding of Dynamic Environments

Difei Gao, Ruiping Wang, Ziyi Bai, Xilin Chen; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 1675-1685

Visual understanding goes well beyond the study of images or videos on the web. To achieve complex tasks in volatile situations, the human can deeply understand the environment, quickly perceive events happening around, and continuously tra ck objects' state changes, which are still challenging for current AI systems. T o equip AI system with the ability to understand dynamic ENVironments, we build a video Question Answering dataset named Env-QA. Env-QA contains 23K egocentric videos, where each video is composed of a series of events about exploring and i nteracting in the environment. It also provides 85K questions to evaluate the ab ility of understanding the composition, layout, and state changes of the environ ment presented by the events in videos. Moreover, we propose a video QA model, T emporal Segmentation and Event Attention network (TSEA), which introduces event- level video representation and corresponding attention mechanisms to better extr act environment information and answer questions. Comprehensive experiments demo nstrate the effectiveness of our framework and show the formidable challenges of Env-QA in terms of long-term state tracking, multi-event temporal reasoning and event counting, etc.

****

Region-Aware Contrastive Learning for Semantic Segmentation

Hanzhe Hu, Jinshi Cui, Liwei Wang; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 16291-16301

Recent works have made great success in semantic segmentation by exploiting cont extual information in a local or global manner within individual image and super vising the model with pixel-wise cross entropy loss. However, from the holistic view of the whole dataset, semantic relations not only exist inside one single i mage, but also prevail in the whole training data, which makes solely considerin g intra-image correlations insufficient. Inspired by recent progress in unsuperv ised contrastive learning, we propose the region-aware contrastive learning (Reg ionContrast) for semantic segmentation in the supervised manner. In order to enh ance the similarity of semantically similar pixels while keeping the discriminat ion from others, we employ contrastive learning to realize this objective. With the help of memory bank, we explore to store all the representative features int o the memory. Without loss of generality, to efficiently incorporate all trainin g data into the memory bank while avoiding taking too much computation resource, we propose to construct region centers to represent features from different cat egories for every image. Hence, the proposed region-aware contrastive learning i s performed in a region level for all the training data, which saves much more m emory than methods exploring the pixel-level relations. The proposed RegionContr ast brings little computation cost during training and requires no extra overhea d for testing. Extensive experiments demonstrate that our method achieves state- of-the-art performance on three benchmark datasets including Cityscapes, ADE20K and COCO Stuff.

****

Image Retrieval on Real-Life Images With Pre-Trained Vision-and-Language Models
Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, Stephen Gould; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2125-2134

We extend the task of composed image retrieval, where an input query consists of an image and short textual description of how to modify the image. Existing methods have only been applied to non-complex images within narrow domains, such as fashion products, thereby limiting the scope of study on in-depth visual reasoning in rich image and language contexts. To address this issue, we collect the Compose Image Retrieval on Real-life images (CIRR) dataset, which consists of over 36,000 pairs of crowd-sourced, open-domain images with human-generated modifying text. To extend current methods to the open-domain, we propose CIRPLANT, a transformer based model that leverages rich pre-trained vision-and-language (V&L) knowledge for modifying visual features conditioned on natural language. Retrieval is then done by nearest neighbor lookup on the modified features. We demonstrate that with a relatively simple architecture, CIRPLANT outperforms existing methods on open-domain images, while matching state-of-the-art accuracy on the existing narrow datasets, such as fashion. Together with the release of CIRR, we believe this work will inspire further research on composed image retrieval. Our dataset, code and pre-trained models are available at https://cuberick-orion.github.io/CIRR/.

****

Self-Supervised Real-to-Sim Scene Generation
Aayush Prakash, Shoubhik Debnath, Jean-Francois Lafleche, Eric Cameracci, Gavriel State, Stan Birchfield, Marc T. Law; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16044-16054

Synthetic data is emerging as a promising solution to the scalability issue of supervised deep learning, especially when real data are difficult to acquire or hard to annotate. Synthetic data generation, however, can itself be prohibitively expensive when domain experts have to manually and painstakingly oversee the process. Moreover, neural networks trained on synthetic data often do not perform well on real data because of the domain gap. To solve these challenges, we propose Sim2SG, a self-supervised automatic scene generation technique for matching the distribution of real data. Importantly, Sim2SG does not require supervision from the real-world dataset, thus making it applicable in situations for which such annotations are difficult to obtain. Sim2SG is designed to bridge both the content and appearance gaps, by matching the content of real data, and by matching the features in the source and target domains. We select scene graph (SG) generation as the downstream task, due to the limited availability of labeled datasets. Experiments demonstrate significant improvements over leading baselines in reducing the domain gap both qualitatively and quantitatively, on several synthetic datasets as well as the real-world KITTI dataset.

****

GP-S3Net: Graph-Based Panoptic Sparse Semantic Segmentation Network
Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, Liu Bingbing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16076-16085

Panoptic segmentation as an integrated task of both static environmental understanding and dynamic object identification, has recently begun to receive broad research interest. In this paper, we propose a new computationally efficient LiDAR based panoptic segmentation framework, called GP-S3Net. GP-S3Net is a proposal-free approach in which no object proposals are needed to identify the objects in contrast to conventional two-stage panoptic systems, where a detection network is incorporated for capturing instance information. Our new design consists of a novel instance-level network to process the semantic results by constructing a graph convolutional network to identify objects (foreground), which later on are fused with the background classes. Through the fine-grained clusters of the foreground objects from the semantic segmentation backbone, over-segmentation priors are generated and subsequently processed by 3D sparse convolution to embed each cluster. Each cluster is treated as a node in the graph and its corresponding

embedding is used as its node feature. Then a GCNN predicts whether edges exist between each cluster pair. We utilize the instance label to generate ground truth edge labels for each constructed graph in order to supervise the learning. Extensive experiments demonstrate that GP-S3Net outperforms the current state-of-the-art approaches, by a significant margin across available datasets such as, nuScenes and SemanticPOSS, ranking first on the competitive public SemanticKITTI leaderboard upon publication.
****

Learning From Noisy Data With Robust Representation Learning
Junnan Li, Caiming Xiong, Steven C.H. Hoi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9485-9494
Learning from noisy data has attracted much attention, where most methods focus on label noise. In this work, we propose a new learning framework which simultaneously addresses three types of noise commonly seen in real-world data: label noise, out-of-distribution input, and input corruption. In contrast to most existing methods, we combat noise by learning robust representation. Specifically, we embed images into a low-dimensional subspace, and regularize the geometric structure of the subspace with robust contrastive learning, which includes an unsupervised consistency loss and a supervised mixup prototypical loss. We also propose a new noise cleaning method which leverages the learned representation to enforce a smoothness constraint on neighboring samples. Experiments on multiple benchmarks demonstrate state-of-the-art performance of our method and robustness of the learned representation. Code is available at https://github.com/salesforce/RRL/.
****

Self-Supervised 3D Skeleton Action Representation Learning With Motion Consistency and Continuity
Yukun Su, Guosheng Lin, Qingyao Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13328-13338
Recently, self-supervised learning (SSL) has been proved very effective and it can help boost the performance in learning representations from unlabeled data in the image domain. Yet, very little is explored about its usefulness in 3D skeleton-based action recognition understanding. Directly applying existing SSL techniques for 3D skeleton learning, however, suffers from trivial solutions and imprecise representations. To tackle these drawbacks, we consider perceiving the consistency and continuity of motion at different playback speeds are two critical issues. To this end, we propose a novel SSL method to learn the 3D skeleton representation in an efficacious way. Specifically, by constructing a positive clip (speed-changed) and a negative clip (motion-broken) of the sampled action sequence, we encourage the positive pairs closer while pushing the negative pairs to force the network to learn the intrinsic dynamic motion consistency information. Moreover, to enhance the learning features, skeleton interpolation is further exploited to model the continuity of human skeleton data. To validate the effectiveness of the proposed method, extensive experiments are conducted on Kinetics, NTU60, NTU120, and PKUMMD datasets with several alternative network architectures. Experimental evaluations demonstrate the superiority of our approach and through which, we can gain significant performance improvement without using extra labeled data.
****

Feature Importance-Aware Transferable Adversarial Attacks
Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, Kui Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7639-7648
Transferability of adversarial examples is of central importance for attacking an unknown model, which facilitates adversarial attacks in more practical scenarios, e.g., blackbox attacks. Existing transferable attacks tend to craft adversarial examples by indiscriminately distorting features to degrade prediction accuracy in a source model without aware of intrinsic features of objects in the images. We argue that such brute-force degradation would introduce model-specific local optimum into adversarial examples, thus limiting the transferability. By con

trast, we propose the Feature Importance-aware Attack (FIA), which disrupts impo rtant object-aware features that dominate model decisions consistently. More spe cifically, we obtain feature importance by introducing the aggregate gradient, w hich averages the gradients with respect to feature maps of the source model, co mputed on a batch of random transforms of the original clean image. The gradient s will be highly correlated to objects of interest, and such correlation present s invariance across different models. Besides, the random transforms will preser ve intrinsic features of objects and suppress model-specific information. Finall y, the feature importance guides to search for adversarial examples towards disr upting critical features, achieving stronger transferability. Extensive experime ntal evaluation demonstrates the effectiveness and superior performance of the p roposed FIA, i.e., improving the success rate by 9.5% against normally trained m odels and 12.8% against defense models as compared to the state-of-the-art trans ferable attacks. Code is available at: https://github.com/hcguoO0/FIA

****

Exploring Classification Equilibrium in Long-Tailed Object Detection

Chengjian Feng, Yujie Zhong, Weilin Huang; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 3417-3426

The conventional detectors tend to make imbalanced classification and suffer per formance drop, when the distribution of the training data is severely skewed. In this paper, we propose to use the mean classification score to indicate the cla ssification accuracy for each category during training. Based on this indicator, we balance the classification via an Equilibrium Loss (EBL) and a Memory-augmen ted Feature Sampling (MFS) method. Specifically, EBL increases the intensity of the adjustment of the decision boundary for the weak classes by a designed score -guided loss margin between any two classes. On the other hand, MFS improves the frequency and accuracy of the adjustments of the decision boundary for the weak classes through over-sampling the instance features of those classes. Therefore , EBL and MFS work collaboratively for finding the classification equilibrium in long-tailed detection, and dramatically improve the performance of tail classes while maintaining or even improving the performance of head classes. We conduct experiments on LVIS using Mask R-CNN with various backbones including ResNet-50 -FPN and ResNet-101-FPN to show the superiority of the proposed method. It impro ves the detection performance of tail classes by 15.6 AP, and outperforms the mo st recent long-tailed object detectors by more than 1 AP. Code is available at h ttps://github.com/fcjian/LOCE.

****

Meta Gradient Adversarial Attack

Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, Shiguang Shan; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 7748-7757

In recent years, research on adversarial attacks has become a hot spot. Although current literature on the transfer-based adversarial attack has achieved promis ing results for improving the transferability to unseen black-box models, it sti ll leaves a long way to go. Inspired by the idea of meta-learning, this paper pr oposes a novel architecture called Meta Gradient Adversarial Attack (MGAA), whic h is plug-and-play and can be integrated with any existing gradient-based attack method for improving the cross-model transferability. Specifically, we randomly sample multiple models from a model zoo to compose different tasks and iterativ ely simulate a white-box attack and a black-box attack in each task. By narrowin g the gap between the gradient directions in white-box and black-box attacks, th e transferability of adversarial examples on the black-box setting can be improv ed. Extensive experiments on the CIFAR10 and ImageNet datasets show that our arc hitecture outperforms the state-of-the-art methods for both black-box and white-box attack settings.

****

Differentiable Convolution Search for Point Cloud Processing

Xing Nie, Yongcheng Liu, Shaohong Chen, Jianlong Chang, Chunlei Huo, Gaofeng Men g, Qi Tian, Weiming Hu, Chunhong Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7437-7446

Exploiting convolutional neural networks for point cloud processing is quite challenging, due to the inherent irregular distribution and discrete shape representation of point clouds. To address these problems, many handcrafted convolution variants have sprung up in recent years. Though with elaborate design, these variants could be far from optimal in sufficiently capturing diverse shapes formed by discrete points. In this paper, we propose PointSeaConv, i.e., a novel differential convolution search paradigm on point clouds. It can work in a purely data-driven manner and thus is capable of auto-creating a group of suitable convolutions for geometric shape modeling. We also propose a joint optimization framework for simultaneous search of internal convolution and external architecture, and introduce epsilon-greedy algorithm to alleviate the effect of discretization error. As a result, PointSeaNet, a deep network that is sufficient to capture geometric shapes at both convolution level and architecture level, can be searched out for point cloud processing. Extensive experiments strongly evidence that our proposed PointSeaNet surpasses current handcrafted deep models on challenging benchmarks across multiple tasks with remarkable margins.

****

Zero-Shot Day-Night Domain Adaptation With a Physics Prior

Attila Lengyel, Sourav Garg, Michael Milford, Jan C. van Gemert; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4399-4409

We explore the zero-shot setting for day-night domain adaptation. The traditional domain adaptation setting is to train on one domain and adapt to the target domain by exploiting unlabeled data samples from the test set. As gathering relevant test data is expensive and sometimes even impossible, we do not rely on test data and instead exploit a visual inductive prior derived from physics-based reflection models for domain adaptation. We cast a number of color invariant edge detectors as trainable layers in a convolutional neural network and evaluate their robustness to illumination changes. We show that the color invariant layer reduces the day-night distribution shift in feature map activations throughout the network. We demonstrate improved performance for zero-shot day to night domain adaptation on both synthetic as well as natural datasets in various tasks, including classification, segmentation and place recognition.

****

Sketch Your Own GAN

Sheng-Yu Wang, David Bau, Jun-Yan Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14050-14060

Can a user create a deep generative model by sketching a single example? Traditionally, creating a GAN model has required the collection of a large-scale dataset of exemplars and specialized knowledge in deep learning. In contrast, sketching is possibly the most universally accessible way to convey a visual concept. In this work, we present a method, GAN Sketching, for rewriting GANs with one or more sketches, to make GANs training easier for novice users. In particular, we change the weights of an original GAN model according to user sketches. We encourage the model's output to match the user sketches through a cross-domain adversarial loss. Furthermore, we explore different regularization methods to preserve the original model's diversity and image quality. Experiments have shown that our method can mold GANs to match shapes and poses specified by sketches while maintaining realism and diversity. Finally, we demonstrate a few applications of the resulting GAN, including latent space interpolation and image editing.

****

Minimal Solutions for Panoramic Stitching Given Gravity Prior

Yaqing Ding, Daniel Barath, Zuzana Kukelova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5579-5588

When capturing panoramas, people tend to align their cameras with the vertical axis, i.e., the direction of gravity. Moreover, modern devices, e.g. smartphones and tablets, are equipped with an IMU (Inertial Measurement Unit) that can measure the gravity vector accurately. Using this prior, the y-axes of the cameras can be aligned or assumed to be already aligned, reducing the relative orientation to 1-DOF (degree of freedom). Exploiting this assumption, we propose new minima

l solutions to panoramic stitching of images taken by cameras with coinciding op
tical centers, i.e. undergoing pure rotation. We consider six practical camera c
onfigurations, from fully calibrated ones up to a camera with unknown fixed or v
arying focal length and with or without radial distortion. The solvers are teste
d both on synthetic scenes, on more than 500k real image pairs from the Sun360 d
ataset, and from scenes captured by us using two smartphones equipped with IMUs.
 The new solvers have similar or better accuracy than the state-of-the-art ones
and outperform them in terms of processing time.
****

iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis
Andreas Blattmann, Timo Milbich, Michael Dorkenwald, Björn Ommer; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1470
7-14717
How would a static scene react to a local poke? What are the effects on other pa
rts of an object if you could locally push it? There will be distinctive movemen
t, despite evident variations caused by the stochastic nature of our world. Thes
e outcomes are governed by the characteristic kinematics of objects that dictate
 their overall motion caused by a local interaction. Conversely, the movement of
 an object provides crucial information about its underlying distinctive kinemat
ics and the interdependencies between its parts. This two-way relation motivates
 learning a bijective mapping between object kinematics and plausible future ima
ge sequences. Therefore, we propose iPOKE -- invertible Prediction of Object Kin
ematics -- that, conditioned on an initial frame and a local poke, allows to sam
ple object kinematics and establishes a one-to-one correspondence to the corresp
onding plausible videos, thereby providing a controlled stochastic video synthes
is. In contrast to previous works, we do not generate arbitrary realistic videos
, but provide efficient control of movements, while still capturing the stochast
ic nature of our environment and the diversity of plausible outcomes it entails.
 Moreover, our approach can transfer kinematics onto novel object instances and
is not confined to particular object classes. Our project page is available at h
ttps://bit.ly/3dJN4Lf.
****

Neural Radiance Flow for 4D View Synthesis and Video Processing
Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, Jiajun Wu; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1
4324-14334
We present a method, Neural Radiance Flow (NeRFlow), to learn a 4D spatial-tempo
ral representation of a dynamic scene from a set of RGB images. Key to our appro
ach is the use of a neural implicit representation that learns to capture the 3D
 occupancy, radiance, and dynamics of the scene. By enforcing consistency across
 different modalities, our representation enables multi-view rendering in divers
e dynamic scenes, including water pouring, robotic interaction, and real images,
 outperforming state-of-the-art methods for spatial-temporal view synthesis. Our
 approach works even when being provided only a single monocular real video. We
further demonstrate that the learned representation can serve as an implicit sce
ne prior, enabling video processing tasks such as image super-resolution and de-
noising without any additional supervision.
****

Assignment-Space-Based Multi-Object Tracking and Segmentation
Anwesa Choudhuri, Girish Chowdhary, Alexander G. Schwing; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13598-13607
Multi-object tracking and segmentation (MOTS) is important for understanding dyn
amic scenes in video data. Existing methods perform well on multi-object detecti
on and segmentation for independent video frames, but tracking of objects over t
ime remains a challenge. MOTS methods formulate tracking locally, i.e., frame-by
-frame, leading to sub-optimal results. Classical global methods on tracking ope
rate directly on object detections, which leads to a combinatorial growth in the
 detection space. In contrast, we formulate a global method for MOTS over the sp
ace of assignments rather than detections: First, we find all top-k assignments
of objects detected and segmented between any two consecutive frames and develop

a structured prediction formulation to score assignment sequences across any number of consecutive frames. We use dynamic programming to find the global optimizer of this formulation in polynomial time. Second, we connect objects which reappear after having been out of view for some time. For this we formulate an assignment problem. On the challenging KITTI-MOTS and MOTSChallenge datasets, this achieves state-of-the-art results among methods which don't use depth data.
****

Vi2CLR: Video and Image for Visual Contrastive Learning of Representation
Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1502-1512
In this paper, we introduce a novel self-supervised visual representation learning method which understands both images and videos in a joint learning fashion. The proposed neural network architecture and objectives are designed to obtain two different Convolutional Neural Networks for solving visual recognition tasks in the domain of videos and images. Our method called Video/Image for Visual Contrastive Learning of Representation(Vi2CLR) uses unlabeled videos to exploit dynamic and static visual cues for self-supervised and instances similarity/dissimilarity learning. Vi2CLR optimization pipeline consists of visual clustering part and representation learning based on groups of similar positive instances within a cluster and negative ones from other clusters and learning visual clusters and their distances. We show how a joint self-supervised visual clustering and instance similarity learning with 2D (image) and 3D (video) CovNet encoders yields such robust and near to supervised learning performance. We extensively evaluate the method on downstream tasks like large scale action recognition and image and object classification on datasets like Kinetics, ImageNet, Pascal VOC'07 and UCF101 and achieve outstanding results compared to state-of-the-art self-supervised methods. To the best of our knowledge, the Vi2CLR is the first of its kind self-supervised neural network to tackle both video and image recognition task simultaneously by only using one source of data.
****

R-MSFM: Recurrent Multi-Scale Feature Modulation for Monocular Depth Estimating
Zhongkai Zhou, Xinnan Fan, Pengfei Shi, Yuanxue Xin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12777-12786
In this paper, we propose Recurrent Multi-Scale Feature Modulation (R-MSFM), a new deep network architecture for self-supervised monocular depth estimation. R-MSFM extracts per-pixel features, builds a multi-scale feature modulation module, and iteratively updates an inverse depth through a parameter-shared decoder at the fixed resolution. This architecture enables our R-MSFM to maintain semantically richer while spatially more precise representations and avoid the error propagation caused by the traditional U-Net-like coarse-to-fine architecture widely used in this domain, resulting in strong generalization and efficient parameter count. Experimental results demonstrate the superiority of our proposed R-MSFM both at model size and inference speed, and show the state-of-the-art results on the KITTI benchmark. Code is available at https://github.com/jsczzzk/R-MSFM
****

Spatially Conditioned Graphs for Detecting Human-Object Interactions
Frederic Z. Zhang, Dylan Campbell, Stephen Gould; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13319-13327
We address the problem of detecting human-object interactions in images using graphical neural networks. Unlike conventional methods, where nodes send scaled but otherwise identical messages to each of their neighbours, we propose to condition messages between pairs of nodes on their spatial relationships, resulting in different messages going to neighbours of the same node. To this end, we explore various ways of applying spatial conditioning under a multi-branch structure. Through extensive experimentation we demonstrate the advantages of spatial conditioning for the computation of the adjacency structure, messages and the refined graph features. In particular, we empirically show that as the quality of the bounding boxes increases, their coarse appearance features contribute relatively less to the disambiguation of interactions compared to the spatial information.

Our method achieves an mAP of 31.33% on HICO-DET and 54.2% on V-COCO, significantly outperforming state-of-the-art on fine-tuned detections.
****

G-DetKD: Towards General Distillation Framework for Object Detectors via Contrastive and Semantic-Guided Feature Imitation
Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, Tong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3591-3600
In this paper, we investigate the knowledge distillation (KD) strategy for object detection and propose an effective framework applicable to both homogeneous and heterogeneous student-teacher pairs. The conventional feature imitation paradigm introduces imitation masks to focus on informative foreground areas while excluding the background noises. However, we find that those methods fail to fully utilize the semantic information in all feature pyramid levels, which leads to inefficiency for knowledge distillation between FPN-based detectors. To this end, we propose a novel semantic-guided feature imitation technique, which automatically performs soft matching between feature pairs across all pyramid levels to provide the optimal guidance to the student. To push the envelop even further, we introduce contrastive distillation to effectively capture the information encoded in the relationship between different feature regions. Finally, we propose a generalized detection KD pipeline, which is capable of distilling both homogeneous and heterogeneous detector pairs. Our method consistently outperforms the existing detection KD techniques, and works when (1) components in the framework are used separately and in conjunction; (2) for both homogeneous and heterogenous student-teacher pairs and (3) on multiple detection benchmarks. With a powerful X101-FasterRCNN-Instaboost detector as the teacher, R50-FasterRCNN reaches 44.0% AP, R50-RetinaNet reaches 43.3% AP and R50-FCOS reaches 43.1% AP on COCO dataset.
****

End-to-End Detection and Pose Estimation of Two Interacting Hands
Dong Uk Kim, Kwang In Kim, Seungryul Baek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11189-11198
Three dimensional hand pose estimation has reached a level of maturity, enabling real-world applications for single-hand cases. However, accurate estimation of the pose of two closely interacting hands still remains a challenge as in this case, one hand often occludes the other. We present a new algorithm that accurately estimates hand poses in such a challenging scenario. The crux of our algorithm lies in a framework that jointly trains the estimators of interacting hands, leveraging their inter-dependence. Further, we employ a GAN-type discriminator of interacting hand pose that helps avoid physically implausible configurations, e.g intersecting fingers, and exploit the visibility of joints to improve intermediate 2D pose estimation. We incorporate them into a single model that learns to detect hands and estimate their pose based on a unified criterion of pose estimation accuracy. To our knowledge, this is the first attempt to build an end-to-end network that detects and estimates the pose of two closely interacting hands (as well as single hands). In the experiments with three datasets representing challenging real-world scenarios, our algorithm demonstrated significant and consistent performance improvements over state-of-the-arts.
****

Fog Simulation on Real LiDAR Point Clouds for 3D Object Detection in Adverse Weather
Martin Hahner, Christos Sakaridis, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15283-15292
This work addresses the challenging task of LiDAR-based 3D object detection in foggy weather. Collecting and annotating data in such a scenario is very time, labor and cost intensive. In this paper, we tackle this problem by simulating physically accurate fog into clear-weather scenes, so that the abundant existing real datasets captured in clear weather can be repurposed for our task. Our contributions are twofold: 1) We develop a physically valid fog simulation method that

is applicable to any LiDAR dataset. This unleashes the acquisition of large-scale foggy training data at no extra cost. These partially synthetic data can be used to improve the robustness of several perception methods, such as 3D object detection and tracking or simultaneous localization and mapping, on real foggy data. 2) Through extensive experiments with several state-of-the-art detection approaches, we show that our fog simulation can be leveraged to significantly improve the performance for 3D object detection in the presence of fog. Thus, we are the first to provide strong 3D object detection baselines on the Seeing Through Fog dataset. Our code is available at www.trace.ethz.ch/lidar_fog_simulation.
****

Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better
Bojia Zi, Shihao Zhao, Xingjun Ma, Yu-Gang Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16443-16452
Adversarial training is one effective approach for training robust deep neural networks against adversarial attacks. While being able to bring reliable robustness, adversarial training (AT) methods in general favor high capacity models, i.e., the larger the model the better the robustness. This tends to limit their effectiveness on small models, which are more preferable in scenarios where storage or computing resources are very limited (e.g., mobile devices). In this paper, we leverage the concept of knowledge distillation to improve the robustness of small models by distilling from adversarially trained large models. We first revisit several state-of-the-art AT methods from a distillation perspective and identify one common technique that can lead to improved robustness: the use of robust soft labels -- predictions of a robust model. Following this observation, we propose a novel adversarial robustness distillation method called Robust Soft Label Adversarial Distillation (RSLAD) to train robust small student models. RSLAD fully exploits the robust soft labels produced by a robust (adversarially-trained) large teacher model to guide the student's learning on both natural and adversarial examples in all loss terms. We empirically demonstrate the effectiveness of our RSLAD approach over existing adversarial training and distillation methods in improving the robustness of small models against state-of-the-art attacks including the AutoAttack. We also provide a set of understandings on our RSLAD and the importance of robust soft labels for adversarial robustness distillation. Code: https://github.com/zibojia/RSLAD.
****

Normalization Matters in Weakly Supervised Object Localization
Jeesoo Kim, Junsuk Choe, Sangdoo Yun, Nojun Kwak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3427-3436
Weakly-supervised object localization (WSOL) enables finding an object using a dataset without any localization information. By simply training a classification model using only image-level annotations, the feature map of a model can be utilized as a score map for localization. In spite of many WSOL methods proposing novel strategies, there has not been any de facto standards about how to normalize the class activation map (CAM). Consequently, many WSOL methods have failed to fully exploit their own capacity because of the misuse of a normalization method. In this paper, we review many existing normalization methods and point out that they should be used according to the property of the given dataset. Additionally, we propose a new normalization method which substantially enhances the performance of any CAM-based WSOL methods. Using the proposed normalization method, we provide a comprehensive evaluation over three datasets (CUB, ImageNet and OpenImages) on three different architectures and observe significant performance gains over the conventional normalization methods in all the evaluated cases.
****

Joint Inductive and Transductive Learning for Video Object Segmentation
Yunyao Mao, Ning Wang, Wengang Zhou, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9670-9679
Semi-supervised video object segmentation is a task of segmenting the target object in a video sequence given only a mask annotation in the first frame. The limited information available makes it an extremely challenging task. Most previous

best-performing methods adopt matching-based transductive reasoning or online inductive learning. Nevertheless, they are either less discriminative for similar instances or insufficient in the utilization of spatio-temporal information. In this work, we propose to integrate transductive and inductive learning into a unified framework to exploit the complementarity between them for accurate and robust video object segmentation. The proposed approach consists of two functional branches. The transduction branch adopts a lightweight transformer architecture to aggregate rich spatio-temporal cues while the induction branch performs online inductive learning to obtain discriminative target information. To bridge these two diverse branches, a two-head label encoder is introduced to learn the suitable target prior for each of them. The generated mask encodings are further forced to be disentangled to better retain their complementarity. Extensive experiments on several prevalent benchmarks show that, without the need of synthetic training data, the proposed approach sets a series of new state-of-the-art records. Code is available at https://github.com/maoyunyao/JOINT.

****

Contrast and Order Representations for Video Self-Supervised Learning

Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, Zhiqiang Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7939-7949

This paper studies the problem of learning self-supervised representations on videos. In contrast to image modality that only requires appearance information on objects or scenes, video needs to further explore the relations between multiple frames/clips along the temporal dimension. However, the recent proposed contrastive-based self-supervised frameworks do not grasp such relations explicitly since they simply utilize two augmented clips from the same video and compare their distance without referring to their temporal relation. To address this, we present a contrast-and-order representation (CORP) framework for learning self-supervised video representations that can automatically capture both the appearance information within each frame and temporal information across different frames. In particular, given two video clips, our model first predicts whether they come from the same input video, and then predict the temporal ordering of the clips if they come from the same video. We also propose a novel decoupling attention method to learn symmetric similarity (contrast) and anti-symmetric patterns (order). Such design involves neither extra parameters nor computation, but can speed up the learning process and improve accuracy compared to the vanilla multi-head attention. We extensively validate the representation ability of our learned video features for the downstream action recognition task on Kinetics-400 and Something-something V2. Our method outperforms previous state-of-the-arts by a significant margin.

****

Out-of-Core Surface Reconstruction via Global TGV Minimization

Nikolai Poliarnyi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5641-5650

We present an out-of-core variational approach for surface reconstruction from a set of aligned depth maps. Input depth maps are supposed to be reconstructed from regular photos or/and can be a representation of terrestrial LIDAR point clouds. Our approach is based on surface reconstruction via total generalized variation minimization (TGV) because of its strong visibility-based noise-filtering properties and GPU-friendliness. Our main contribution is an out-of-core OpenCL-accelerated adaptation of this numerical algorithm which can handle arbitrarily large real-world scenes with scale diversity.

****

Skeleton Cloud Colorization for Unsupervised 3D Action Representation Learning

Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, Alex C. Kot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13423-13433

Skeleton-based human action recognition has attracted increasing attention in recent years. However, most of the existing works focus on supervised learning which requiring a large number of annotated action sequences that are often expensi

ve to collect. We investigate unsupervised representation learning for skeleton action recognition, and design a novel skeleton cloud colorization technique that is capable of learning skeleton representations from unlabeled skeleton sequence data. Specifically, we represent a skeleton action sequence as a 3D skeleton cloud and colorize each point in the cloud according to its temporal and spatial orders in the original (unannotated) skeleton sequence. Leveraging the colorized skeleton point cloud, we design an auto-encoder framework that can learn spatial-temporal features from the artificial color labels of skeleton joints effectively. We evaluate our skeleton cloud colorization approach with action classifiers trained under different configurations, including unsupervised, semi-supervised and fully-supervised settings. Extensive experiments on NTU RGB+D and NW-UCLA datasets show that the proposed method outperforms existing unsupervised and semi-supervised 3D action recognition methods by large margins, and it achieves competitive performance in supervised 3D action recognition as well.

****

Latent Transformations via NeuralODEs for GAN-Based Image Editing

Valentin Khrulkov, Leyla Mirvakhabova, Ivan Oseledets, Artem Babenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14428-14437

Recent advances in high-fidelity semantic image editing heavily rely on the presumably disentangled latent spaces of the state-of-the-art generative models, such as StyleGAN. Specifically, recent works show that it is possible to achieve decent controllability of attributes in the face images via linear shifts along with latent directions. Several recent methods address the discovery of such directions, implicitly assuming that the state-of-the-art GANs learn the latent spaces with inherently linearly separable attribute distributions and semantic vector arithmetic properties. In our work, we show that nonlinear latent code manipulations realized as flows of a trainable Neural ODE are beneficial for many practical non-face image domains with more complex non-textured factors of variation. In particular, we investigate a large number of datasets with known attributes and demonstrate that certain attribute manipulations are challenging to be obtained with linear shifts only.

****

DECA: Deep Viewpoint-Equivariant Human Pose Estimation Using Capsule Autoencoders

Nicola Garau, Niccolò Bisagno, Piotr Bródka, Nicola Conci; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11677-11686

Human Pose Estimation (HPE) aims at retrieving the 3D position of human joints from images or videos. We show that current 3D HPE methods suffer a lack of viewpoint equivariance, namely they tend to fail or perform poorly when dealing with viewpoints unseen at training time. Deep learning methods often rely on either scale-invariant, translation-invariant, or rotation-invariant operations, such as max-pooling. However, the adoption of such procedures does not necessarily improve viewpoint generalization, rather leading to more data-dependent methods. To tackle this issue, we propose a novel capsule autoencoder network with fast Variational Bayes capsule routing, named DECA. By modeling each joint as a capsule entity, combined with the routing algorithm, our approach can preserve the joints' hierarchical and geometrical structure in the feature space, independently from the viewpoint. By achieving viewpoint equivariance, we drastically reduce the network data dependency at training time, resulting in an improved ability to generalize for unseen viewpoints. In the experimental validation, we outperform other methods on depth images from both seen and unseen viewpoints, both top-view, and front-view. In the RGB domain, the same network gives state-of-the-art results on the challenging viewpoint transfer task, also establishing a new framework for top-view HPE.

****

Extensions of Karger's Algorithm: Why They Fail in Theory and How They Are Useful in Practice

Erik Jenner, Enrique Fita Sanmartín, Fred A. Hamprecht; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4602-4611

The minimum graph cut and minimum s-t-cut problems are important primitives in t
he modeling of combinatorial problems in computer science, including in computer
 vision and machine learning. Some of the most efficient algorithms for finding
global minimum cuts are randomized algorithms based on Karger's groundbreaking c
ontraction algorithm. Here, we study whether Karger's algorithm can be successfu
lly generalized to other cut problems. We first prove that a wide class of natur
al generalizations of Karger's algorithm cannot efficiently solve the s-t-mincut
 or the normalized cut problem to optimality. However, we then present a simple
new algorithm for seeded segmentation / graph-based semi-supervised learning tha
t is closely based on Karger's original algorithm, showing that for these proble
ms, extensions of Karger's algorithm can be useful. The new algorithm has linear
 asymptotic runtime and yields a potential that can be interpreted as the poster
ior probability of a sample belonging to a given seed / class. We clarify its re
lation to the random walker algorithm / harmonic energy minimization in terms of
 distributions over spanning forests. On classical problems from seeded image se
gmentation and graph-based semi-supervised learning on image data, the method pe
rforms at least as well as the random walker / harmonic energy minimization / Ga
ussian processes.
****

Geometry-Free View Synthesis: Transformers and No 3D Priors
Robin Rombach, Patrick Esser, Björn Ommer; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 14356-14366
Is a geometric model required to synthesize novel views from a single image? Bei
ng bound to local convolutions, CNNs need explicit 3D biases to model geometric
transformations. In contrast, we demonstrate that a transformer-based model can
synthesize entirely novel views without any hand-engineered 3D biases. This is a
chieved by (i) a global attention mechanism for implicitly learning long-range 3
D correspondences between source and target views, and (ii) a probabilistic form
ulation necessary to capture the ambiguity inherent in predicting novel views fr
om a single image, thereby overcoming the limitations of previous approaches tha
t are restricted to relatively small viewpoint changes. We evaluate various ways
 to integrate 3D priors into a transformer architecture. However, our experiment
s show that no such geometric priors are required and that the transformer is ca
pable of implicitly learning 3D relationships between images. Furthermore, this
approach outperforms the state of the art in terms of visual quality while cover
ing the full distribution of possible realizations.
****

Scaling-Up Disentanglement for Image Translation
Aviv Gabbay, Yedid Hoshen; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2021, pp. 6783-6792
Image translation methods typically aim to manipulate a set of labeled attribute
s (given as supervision at training time e.g. domain label) while leaving the un
labeled attributes intact. Current methods achieve either: (i) disentanglement,
which exhibits low visual fidelity and can only be satisfied where the attribute
s are perfectly uncorrelated. (ii) visually-plausible translations, which are cl
early not disentangled. In this work, we propose OverLORD, a single framework fo
r disentangling labeled and unlabeled attributes as well as synthesizing high-fi
delity images, which is composed of two stages; (i) Disentanglement: Learning di
sentangled representations with latent optimization. Differently from previous a
pproaches, we do not rely on adversarial training or any architectural biases. (
ii) Synthesis: Training feed-forward encoders for inferring the learned attribut
es and tuning the generator in an adversarial manner to increase the perceptual
quality. When the labeled and unlabeled attributes are correlated, we model an a
dditional representation that accounts for the correlated attributes and improve
s disentanglement. We highlight that our flexible framework covers multiple sett
ings as disentangling labeled attributes, pose and appearance, localized concept
s, and shape and texture. We present significantly better disentanglement with h
igher translation quality and greater output diversity than state-of-the-art met
hods.
****

MeshTalk: 3D Face Animation From Speech Using Cross-Modality Disentanglement
Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, Yaser Sheikh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1173-1182
This paper presents a generic method for generating full facial 3D animation from speech. Existing approaches to audio-driven facial animation exhibit uncanny or static upper face animation, fail to produce accurate and plausible co-articulation or rely on person-specific models that limit their scalability. To improve upon existing models, we propose a generic audio-driven facial animation approach that achieves highly realistic motion synthesis results for the entire face. At the core of our approach is a categorical latent space for facial animation that disentangles audio-correlated and audio-uncorrelated information based on a novel cross-modality loss. Our approach ensures highly accurate lip motion, while also synthesizing plausible animation of the parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. We demonstrate that our approach outperforms several baselines and obtains state-of-the-art quality both qualitatively and quantitatively. A perceptual user study demonstrates that our approach is deemed more realistic than the current state-of-the-art in over 75% of cases. We recommend watching the supplemental video before reading the paper: https://github.com/facebookresearch/meshtalk
****
Learning a Single Network for Scale-Arbitrary Super-Resolution
Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, Yulan Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4801-4810
Recently, the performance of single image super-resolution (SR) has been significantly improved with powerful networks. However, these networks are developed for image SR with specific integer scale factors (e.g., x2/3/4), and cannot handle non-integer and asymmetric SR. In this paper, we propose to learn a scale-arbitrary image SR network from scale-specific networks. Specifically, we develop a plug-in module for existing SR networks to perform scale-arbitrary SR, which consists of multiple scale-aware feature adaption blocks and a scale-aware upsampling layer. Moreover, conditional convolution is used in our plug-in module to generate dynamic scale-aware filters, which enables our network to adapt to arbitrary scale factors. Our plug-in module can be easily adapted to existing networks to realize scale-arbitrary SR with a single model. These networks plugged with our module can produce promising results for non-integer and asymmetric SR while maintaining state-of-the-art performance for SR with integer scale factors. Besides, the additional computational and memory cost of our module is very small.
****
Salient Object Ranking With Position-Preserved Attention
Hao Fang, Daoxin Zhang, Yi Zhang, Minghao Chen, Jiawei Li, Yao Hu, Deng Cai, Xiaofei He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16331-16341
Instance segmentation can detect where the objects are in an image, but hard to understand the relationship between them. We pay attention to a typical relationship, relative saliency. A closely related task, salient object detection, predicts a binary map highlighting a visually salient region while hard to distinguish multiple objects. Directly combining two tasks by post-processing also leads to poor performance. There is a lack of research on relative saliency at present, limiting the practical applications such as content-aware image cropping, video summary, and image labeling. In this paper, we study the Salient Object Ranking (SOR) task, which manages to assign a ranking order of each detected object according to its visual saliency. We propose the first end-to-end framework of the SOR task and solve it in a multi-task learning fashion. The framework handles instance segmentation and salient object ranking simultaneously. In this framework, the SOR branch is independent and flexible to cooperate with different detection methods, so that easy to use as a plugin. We also introduce a Position-Preserved Attention (PPA) module tailored for the SOR branch. It consists of the position embedding stage and feature interaction stage. Considering the importance of

position in saliency comparison, we preserve absolute coordinates of objects in ROI pooling operation and then fuse positional information with semantic features in the first stage. In the feature interaction stage, we apply the attention mechanism to obtain proposals' contextualized representations to predict their relative ranking orders. Extensive experiments have been conducted on the ASR dataset. Without bells and whistles, our proposed method outperforms the former state-of-the-art method significantly. The code will be released publicly available on https://github.com/EricFH/SOR.

****

Paint Transformer: Feed Forward Neural Painting With Stroke Prediction

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, Hao Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6598-6607

Neural painting refers to the procedure of producing a series of strokes for a given image and non-photo-realistically recreating it using neural networks. While reinforcement learning (RL) based agents can generate a stroke sequence step by step for this task, it is not easy to train a stable RL agent. On the other hand, stroke optimization methods search for a set of stroke parameters iteratively in a large search space; such low efficiency significantly limits their prevalence and practicality. Different from previous methods, in this paper, we formulate the task as a set prediction problem and propose a novel Transformer-based framework, dubbed Paint Transformer, to predict the parameters of a stroke set with a feed forward network. This way, our model can generate a set of strokes in parallel and obtain the final painting of size 512x512 in near real time. More importantly, since there is no dataset available for training the Paint Transformer, we devise a self-training pipeline such that it can be trained without any off-the-shelf dataset while still achieving excellent generalization capability. Experiments demonstrate that our method achieves better painting performance than previous ones with cheaper training and inference costs. Codes and models will be available.

****

DetCo: Unsupervised Contrastive Learning for Object Detection

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8392-8401

We present DetCo, a simple yet effective self-supervised approach for object detection. Unsupervised pre-training methods have been recently designed for object detection, but they are usually deficient in image classification, or the opposite. Unlike them, DetCo transfers well on downstream instance-level dense prediction tasks, while maintaining competitive image-level classification accuracy. The advantages are derived from (1) multi-level supervision to intermediate representations, (2) contrastive learning between global image and local patches. These two designs facilitate discriminative and consistent global and local representation at each level of feature pyramid, improving detection and classification, simultaneously. Extensive experiments on VOC, COCO, Cityscapes, and ImageNet demonstrate that DetCo not only outperforms recent methods on a series of 2D and 3D instance-level detection tasks, but also competitive on image classification. For example, on ImageNet classification, DetCo is 6.9% and 5.0% top-1 accuracy better than InsLoc and DenseCL, which are two contemporary works designed for object detection. Moreover, on COCO detection, DetCo is 6.9 AP better than SwAV with Mask R-CNN C4. Notably, DetCo largely boosts up Sparse R-CNN, a recent strong detector, from 45.0 AP to 46.5 AP (+1.5 AP), establishing a new SOTA on COCO. Code is available.

****

PR-RRN: Pairwise-Regularized Residual-Recursive Networks for Non-Rigid Structure-From-Motion

Haitian Zeng, Yuchao Dai, Xin Yu, Xiaohan Wang, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5600-5609

We propose PR-RRN, a novel neural-network based method for Non-rigid Structure-from-Motion (NRSfM). PR-RRN consists of Residual-Recursive Networks (RRN) and two

extra regularization losses. RRN is designed to effectively recover 3D shape and camera from 2D keypoints with novel residual-recursive structure. As NRSfM is a highly under-constrained problem, we propose two new pairwise regularization to further regularize the reconstruction. The Rigidity-based Pairwise Contrastive Loss regularizes the shape representation by encouraging higher similarity between the representations of high-rigidity pairs of frames than low-rigidity pairs. We propose minimum singular-value ratio to measure the pairwise rigidity. The Pairwise Consistency Loss enforces the reconstruction to be consistent when the estimated shapes and cameras are exchanged between pairs. Our approach achieves state-of-the-art performance on CMU MOCAP and PASCAL3D+ dataset.
****

YouRefIt: Embodied Reference Understanding With Language and Gesture
Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, Siyuan Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1385-1395
We study the machine's understanding of embodied reference: One agent uses both language and gesture to refer to an object to another agent in a shared physical environment. Of note, this new visual task requires understanding multimodal cues with perspective-taking to identify which object is being referred to. To tackle this problem, we introduce YouRefIt, a new crowd-sourced dataset of embodied reference collected in various physical scenes; the dataset contains 4,195 unique reference clips in 432 indoor scenes. To the best of our knowledge, this is the first embodied reference dataset that allows us to study referring expressions in daily physical scenes to understand referential behavior, human communication, and human-robot interaction. We further devise two benchmarks for image-based and video-based embodied reference understanding. Comprehensive baselines and extensive experiments provide the very first result of machine perception on how the referring expressions and gestures affect the embodied reference understanding. Our results provide essential evidence that gestural cues are as critical as language cues in understanding the embodied reference.
****

Omniscient Video Super-Resolution
Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, Jiayi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4429-4438
Most recent video super-resolution (SR) methods either adopt an iterative manner to deal with low-resolution (LR) frames from a temporally sliding window, or leverage the previously estimated SR output to help reconstruct the current frame recurrently. A few studies try to combine these two structures to form a hybrid framework but have failed to give full play to it. In this paper, we propose an omniscient framework to not only utilize the preceding SR output, but also leverage the SR outputs from the present and future. The omniscient framework is more generic because the iterative, recurrent and hybrid frameworks can be regarded as its special cases. The proposed omniscient framework enables a generator to behave better than its counterparts under other frameworks. Abundant experiments on public datasets show that our method is superior to the state-of-the-art methods in objective metrics, subjective visual effects and complexity.
****

Clothing Status Awareness for Long-Term Person Re-Identification
Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, ZhaoXiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11895-11904
Long-Term person re-identification (LT-reID) exposes extreme challenges because of the longer time gaps between two recording footages where a person is likely to change clothing. There are two types of approaches for LT-reID: biometrics-based approach and data adaptation based approach. The former one is to seek clothing irrelevant biometric features. However, seeking high quality biometric feature is the main concern. The latter one adopts fine-tuning strategy by using data with significant clothing change. However, the performance is compromised when it is applied to cases without clothing change. This work argues that these appr

oaches in fact are not aware of clothing status (i.e., change or no-change) of a pedestrian. Instead, they blindly assume all footages of a pedestrian have different clothes. To tackle this issue, a Regularization via Clothing Status Awareness Network (RCSANet) is proposed to regularize descriptions of a pedestrian by embedding the clothing status awareness. Consequently, the description can be enhanced to maintain the best ID discriminative feature while improving its robustness to real-world LT-reID where both clothing-change case and no-clothing-change case exist. Experiments show that RCSANet performs reasonably well on three LT-reID datasets.

****

Exploring Temporal Coherence for More General Video Face Forgery Detection
Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, Fang Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15044-15054

Although current face manipulation techniques achieve impressive performance regarding quality and controllability, they are struggling to generate temporal coherent face videos. In this work, we explore to take full advantage of the temporal coherence for video face forgery detection. To achieve this, we propose a novel end-to-end framework, which consists of two major stages. The first stage is a fully temporal convolution network (FTCN). The key insight of FTCN is to reduce the spatial convolution kernel size to 1, while maintaining the temporal convolution kernel size unchanged. We surprisingly find this special design can benefit the model for extracting the temporal features as well as improve the generalization capability. The second stage is a Temporal Transformer network, which aims to explore the long-term temporal coherence. The proposed framework is general and flexible, which can be directly trained from scratch without any pre-training models or external datasets. Extensive experiments show that our framework outperforms existing methods and remains effective when applied to detect new sorts of face forgery videos.

****

A Lazy Approach to Long-Horizon Gradient-Based Meta-Learning
Muhammad Abdullah Jamal, Liqiang Wang, Boqing Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6577-6586

Gradient-based meta-learning relates task-specific models to a meta-model by gradients. By this design, an algorithm first optimizes the task-specific models by an inner loop and then backpropagates meta-gradients through the loop to update the meta-model. The number of inner-loop optimization steps has to be small (e.g., one step) to avoid high-order derivatives, big memory footprints, and the risk of vanishing or exploding meta-gradients. We propose an intuitive teacher-student scheme to enable the gradient-based meta-learning algorithms to explore long horizons by the inner loop. The key idea is to employ a student network to adequately explore the search space of task-specific models (e.g., by more than ten steps), and a teacher then takes a "leap" toward the regions probed by the student. The teacher not only arrives at a high-quality model but also defines a lightweight computation graph for meta-gradients. Our approach is generic; it performs well when applied to four meta-learning algorithms over three tasks: few-shot learning, long-tailed classification, and meta-attack.

****

Representative Color Transform for Image Enhancement
Hanul Kim, Su-Min Choi, Chang-Su Kim, Yeong Jun Koh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4459-4468

Recently, the encoder-decoder and intensity transformation approaches lead to impressive progress in image enhancement. However, the encoder-decoder often loses details in input images during down-sampling and up-sampling processes. Also, the intensity transformation has a limited capacity to cover color transformation between low-quality and high-quality images. In this paper, we propose a novel approach, called representative color transform (RCT), to tackle these issues in existing methods. RCT determines different representative colors specialized in input images and estimates transformed colors for the representative colors. It then determines enhanced colors using these transformed colors based on the sim

ilarity between input and representative colors. Extensive experiments demonstra te that the proposed algorithm outperforms recent state-of-the-art algorithms on various image enhancement problems.
****

Image Synthesis via Semantic Composition
Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, Jiaya Jia; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13749-13758
In this paper, we present a novel approach to synthesize realistic images based on their semantic layouts. It hypothesizes that for objects with similar appeara nce, they share similar representation. Our method establishes dependencies betw een regions according to their appearance correlation, yielding both spatially v ariant and associated representations. Conditioning on these features, we propos e a dynamic weighted network constructed by spatially conditional computation (w ith both convolution and normalization). More than preserving semantic distincti ons, the given dynamic network strengthens semantic relevance, benefiting global structure and detail synthesis. We demonstrate that our method gives the compel ling generation performance qualitatively and quantitatively with extensive expe riments on benchmarks.
****

Temporally-Coherent Surface Reconstruction via Metric-Consistent Atlases
Jan Bednarik, Vladimir G. Kim, Siddhartha Chaudhuri, Shaifali Parashar, Mathieu Salzmann, Pascal Fua, Noam Aigerman; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2021, pp. 10458-10467
We propose a method for the unsupervised reconstruction of a temporally-coherent sequence of surfaces from a sequence of time-evolving point clouds, yielding de nse, semantically meaningful correspondences between all keyframes. We represent the reconstructed surface as an atlas, using a neural network. Using canonical correspondences defined via the atlas, we encourage the reconstruction to be as isometric as possible across frames, leading to semantically-meaningful reconstr uction. Through experiments and comparisons, we empirically show that our method achieves results that exceed that state of the art in the accuracy of correspon dences and accuracy of surface reconstruction.
****

Online Continual Learning With Natural Distribution Shifts: An Empirical Study W ith Visual Data
Zhipeng Cai, Ozan Sener, Vladlen Koltun; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2021, pp. 8281-8290
Continual learning is the problem of learning and retaining knowledge through ti me over multiple tasks and environments. Research has primarily focused on the i ncremental classification setting, where new tasks/classes are added at discrete time intervals. Such an "offline" setting does not evaluate the ability of agen ts to learn effectively and efficiently, since an agent can perform multiple lea rning epochs without any time limitation when a task is added. We argue that "on line" continual learning, where data is a single continuous stream without task boundaries, enables evaluating both information retention and online learning ef ficacy. In online continual learning, each incoming small batch of data is first used for testing and then added to the training set, making the problem truly o nline. Trained models are later evaluated on historical data to assess informati on retention. We introduce a new benchmark for online continual visual learning that exhibits large scale and natural distribution shifts. Through a large-scale analysis, we identify critical and previously unobserved phenomena of gradient-based optimization in continual learning, and propose effective strategies for i mproving gradient-based online continual learning with real data. The source cod e and dataset are available in: https://github.com/ IntelLabs/continuallearning.
****

Social Fabric: Tubelet Compositions for Video Relation Detection
Shuo Chen, Zenglin Shi, Pascal Mettes, Cees G. M. Snoek; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 13485-13494
This paper strives to classify and detect the relationship between object tubele ts appearing within a video as a <subject-predicate-object> triplet. Where exist

ing works treat object proposals or tubelets as single entities and model their relations a posteriori, we propose to classify and detect predicates for pairs of object tubelets a priori. We also propose Social Fabric: an encoding that represents a pair of object tubelets as a composition of interaction primitives. These primitives are learned over all relations, resulting in a compact representation able to localize and classify relations from the pool of co-occurring object tubelets across all timespans in a video. The encoding enables our two-stage network. In the first stage, we train Social Fabric to suggest proposals that are likely interacting. We use the Social Fabric in the second stage to simultaneously fine-tune and predict predicate labels for the tubelets. Experiments demonstrate the benefit of early video relation modeling, our encoding and the two-stage architecture, leading to a new state-of-the-art on two benchmarks. We also show how the encoding enables query-by-primitive-example to search for spatio-temporal video relations. Code: https://github.com/shanshuo/Social-Fabric.

****

On Feature Decorrelation in Self-Supervised Learning

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, Hang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9598-9608

In self-supervised representation learning, a common idea behind most of the state-of-the-art approaches is to enforce the robustness of the representations to predefined augmentations. A potential issue of this idea is the existence of completely collapsed solutions (i.e., constant features), which are typically avoided implicitly by carefully chosen implementation details. In this work, we study a relatively concise framework containing the most common components from recent approaches. We verify the existence of complete collapse and discover another reachable collapse pattern that is usually overlooked, namely dimensional collapse. We connect dimensional collapse with strong correlations between axes and consider such connection as a strong motivation for feature decorrelation (i.e., standardizing the covariance matrix). The gains from feature decorrelation are verified empirically to highlight the importance and the potential of this insight.

****

Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, Jianfeng Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2998-3008

This paper presents a new Vision Transformer (ViT) architecture Multi-Scale Vision Longformer, which significantly enhances the ViT of [??] for encoding high-resolution images using two techniques. The first is the multi-scale model structure, which provides image encodings at multiple scales with manageable computational cost. The second is the attention mechanism of Vision Longformer, which is a variant of Longformer [??], originally developed for natural language processing, and achieves a linear complexity w.r.t. the number of input tokens. A comprehensive empirical study shows that the new ViT significantly outperforms several strong baselines, including the existing ViT models and their ResNet counterparts, and the Pyramid Vision Transformer from a concurrent work [??], on a range of vision tasks, including image classification, object detection, and segmentation. The models and source code are released at https://github.com/microsoft/vision-longformer.

****

Viewing Graph Solvability via Cycle Consistency

Federica Arrigoni, Andrea Fusiello, Elisa Ricci, Tomas Pajdla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5540-5549

In structure-from-motion the viewing graph is a graph where vertices correspond to cameras and edges represent fundamental matrices. We provide a new formulation and an algorithm for establishing whether a viewing graph is solvable, i.e. it uniquely determines a set of projective cameras. Known theoretical conditions e

ither do not fully characterize the solvability of all viewing graphs, or are ex ceedingly hard to compute for they involve solving a system of polynomial equati ons with a large number of unknowns. The main result of this paper is a method f or reducing the number of unknowns by exploiting the cycle consistency. We advan ce the understanding of the solvability by (i) finishing the classification of a ll previously undecided minimal graphs up to 9 nodes, (ii) extending the practic al solvability testing up to minimal graphs with up to 90 nodes, and (iii) defin itely answering an open research question by showing that the finite solvability is not equivalent to the solvability. Finally, we present an experiment on real data showing that unsolvable graphs are appearing in practical situations.
****

Low Curvature Activations Reduce Overfitting in Adversarial Training
Vasu Singla, Sahil Singla, Soheil Feizi, David Jacobs; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 16423-16433
Adversarial training is one of the most effective defenses against adversarial a ttacks. Previous works suggest that overfitting is a dominant phenomenon in adve rsarial training leading to a large generalization gap between test and train ac curacy in neural networks. In this work, we show that the observed generalizatio n gap is closely related to the choice of the activation function. In particular , we show that using activation functions with low (exact or approximate) curvat ure values has a regularization effect that significantly reduces both the stand ard and robust generalization gaps in adversarial training. We observe this effe ct for both differentiable/smooth activations such as SiLU as well as non-diff erentiable/non-smooth activations such as LeakyReLU. In the latter case, the "ap proximate" curvature of the activation is low. Finally, we show that for activat ion functions with low curvature, the double descent phenomenon for adversariall y trained models does not occur.
****

SACoD: Sensor Algorithm Co-Design Towards Efficient CNN-Powered Intelligent Phla tCam
Yonggan Fu, Yang Zhang, Yue Wang, Zhihan Lu, Vivek Boominathan, Ashok Veeragha van, Yingyan Lin; Proceedings of the IEEE/CVF International Conference on Comput er Vision (ICCV), 2021, pp. 5168-5177
There has been a booming demand for integrating Convolutional Neural Networks (C NNs) powered functionalities into Internet-of-Thing (IoT) devices to enable ubiq uitous intelligent "IoT cameras". However, more extensive applications of such I oT systems are still limited by two challenges. First, some applications, especi ally medicine- and wearable-related ones, impose stringent requirements on the c amera form factor. Second, powerful CNNs often require considerable storage and energy cost, whereas IoT devices often suffer from limited resources. PhlatCam, with its form factor potentially reduced by orders of magnitude, has emerged as a promising solution to the first aforementioned challenge, while the second one remains a bottleneck. Existing compression techniques, which can potentially ta ckle the second challenge, are far from realizing the full potential in storage and energy reduction, because they mostly focus on the CNN algorithm itself. To this end, this work proposes SACoD, a Sensor Algorithm Co-Design framework to de velop more efficient CNN-powered PhlatCam. In particular, the mask coded in the PhlatCam sensor and the backend CNN model are jointly optimized in terms of both model parameters and architectures via differential neural architecture search. Extensive experiments including both simulation and physical measurement on man ufactured masks show that the proposed SACoD framework achieves aggressive model compression and energy savings while maintaining or even boosting the task accu racy, when benchmarking over two state-of-the-art (SOTA) designs with six datase ts across four different vision tasks including classification, segmentation, im age translation, and face recognition. Our codes are available at: https://githu b.com/RICE-EIC/SACoD.
****

Adaptive Focus for Efficient Video Recognition
Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, Gao Huang; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,

pp. 16249-16258
In this paper, we explore the spatial redundancy in video recognition with the a im to improve the computational efficiency. It is observed that the most informa tive region in each frame of a video is usually a small image patch, which shift s smoothly across frames. Therefore, we model the patch localization problem as a sequential decision task, and propose a reinforcement learning based approach for efficient spatially adaptive video recognition (AdaFocus). In specific, a li ght-weighted ConvNet is first adopted to quickly process the full video sequence , whose features are used by a recurrent policy network to localize the most tas k-relevant regions. Then the selected patches are inferred by a high-capacity ne twork for the final prediction. During offline inference, once the informative p atch sequence has been generated, the bulk of computation can be done in paralle l, and is efficient on modern GPU devices. In addition, we demonstrate that the proposed method can be easily extended by further considering the temporal redun dancy, e.g., dynamically skipping less valuable frames. Extensive experiments on five benchmark datasets, i.e., ActivityNet, FCVID, Mini-Kinetics, Something-Som ething V1&V2, demonstrate that our method is significantly more efficient than t he competitive baselines. Code is available at https://github. com/blackfeather-wang/AdaFocus.
****

Audio2Gestures: Generating Diverse Gestures From Speech Audio With Conditional V ariational Autoencoders
Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, Linchao Bao; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 11293-11302
Generating conversational gestures from speech audio is challenging due to the i nherent one-to-many mapping between audio and body motions. Conventional CNNs/RN Ns assume one-to-one mapping, and thus tend to predict the average of all possib le target motions, resulting in plain/boring motions during inference. In order to overcome this problem, we propose a novel conditional variational autoencoder (VAE) that explicitly models one-to-many audio-to-motion mapping by splitting t he cross-modal latent code into shared code and motion-specific code. The shared code mainly models the strong correlation between audio and motion (such as the synchronized audio and motion beats), while the motion-specific code captures d iverse motion information independent of the audio. However, splitting the laten t code into two parts poses training difficulties for the VAE model. A mapping n etwork facilitating random sampling along with other techniques including relaxe d motion loss, bicycle constraint, and diversity loss are designed to better tra in the VAE. Experiments on both 3D and 2D motion datasets verify that our method generates more realistic and diverse motions than state-of-the-art methods, qua ntitatively and qualitatively. Finally, we demonstrate that our method can be re adily used to generate motion sequences with user-specified motion clips on the timeline. Code and more results are at https://jingli513.github.io/audio2gesture s.
****

SnowflakeNet: Point Cloud Completion by Snowflake Point Deconvolution With Skip-Transformer
Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, Zhizhong Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 5499-5509
Point cloud completion aims to predict a complete shape in high accuracy from it s partial observation. However, previous methods usually suffered from discrete nature of point cloud and unstructured prediction of points in local regions, wh ich makes it hard to reveal fine local geometric details on the complete shape. To resolve this issue, we propose SnowflakeNet with Snowflake Point Deconvolutio n (SPD) to generate the complete point clouds. The SnowflakeNet models the gener ation of complete point clouds as the snowflake-like growth of points in 3D spac e, where the child points are progressively generated by splitting their parent points after each SPD. Our insight of revealing detailed geometry is to introduc e skip-transformer in SPD to learn point splitting patterns which can fit local

regions the best. Skip-transformer leverages attention mechanism to summarize th
e splitting patterns used in the previous SPD layer to produce the splitting in
the current SPD layer. The locally compact and structured point cloud generated
by SPD is able to precisely capture the structure characteristic of 3D shape in
local patches, which enables the network to predict highly detailed geometries,
such as smooth regions, sharp edges and corners. Our experimental results outper
form the state-of-the-art point cloud completion methods under widely used bench
marks. Code will be available at https://github.com/AllenXiangX/SnowflakeNet.
****

LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference
Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin,
Hervé Jégou, Matthijs Douze; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 12259-12269
We design a family of image classification architectures that optimize the trade
-off between accuracy and efficiency in a high-speed regime. Our work exploits r
ecent findings in attention-based architectures, which are competitive on highly
 parallel processing hardware. We revisit principles from the extensive literatu
re on convolutional neural networks to apply them to transformers, in particular
 activation maps with decreasing resolutions. We also introduce the attention bi
as, a new way to integrate positional information in vision transformers. As a r
esult, we propose LeViT: a hybrid neural network for fast inference image classi
fication. We consider different measures of efficiency on different hardware pla
tforms, so as to best reflect a wide range of application scenarios. Our extensi
ve experiments empirically validate our technical choices and show they are suit
able to most architectures. Overall, LeViT significantly outperforms existing co
nvnets and vision transformers with respect to the speed/accuracy tradeoff. For
example, at 80% ImageNet top-1 accuracy, LeViT is 5 times faster than EfficientN
et on CPU. We release the code at https://github.com/facebookresearch/LeViT.
****

Active Universal Domain Adaptation
Xinhong Ma, Junyu Gao, Changsheng Xu; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2021, pp. 8968-8977
Most unsupervised domain adaptation methods rely on rich prior knowledge about t
he source-target label set relationship, and they cannot recognize categories be
yond the source classes, which limits their applicability in practical scenarios
. This paper proposes a new paradigm for unsupervised domain adaptation, termed
as Active Universal Domain Adaptation (AUDA), which removes all label set assump
tions and aims for not only recognizing target samples from source classes but a
lso inferring those from target-private classes by using active learning to anno
tate a small budget of target data. For AUDA, it is challenging to jointly adapt
 the model to the target domain and select informative target samples for annota
tions under a large domain gap and significant semantic shift. To address the pr
oblems, we propose an Active Universal Adaptation Network (AUAN). Specifically,
we first introduce Adversarial and Diverse Curriculum Learning (ADCL), which pro
gressively aligns source and target domains to classify whether target samples a
re from source classes. Then, we propose a Clustering Non-transferable Gradient
Embedding (CNTGE) strategy, which utilizes the clues of transferability, diversi
ty, and uncertainty to annotate target informative sample, making it possible to
 infer labels for target samples of target-private classes. Finally, we propose
to jointly train ADCL and CNTGE with target supervision to promote domain adapta
tion and target-private class recognition. Extensive experiments demonstrate tha
t the proposed AUDA model equipped with ADCL and CNTGE achieves significant resu
lts on four popular benchmarks.
****

FairNAS: Rethinking Evaluation Fairness of Weight Sharing Neural Architecture Se
arch
Xiangxiang Chu, Bo Zhang, Ruijun Xu; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2021, pp. 12239-12248
One of the most critical problems in weight-sharing neural architecture search i
s the evaluation of candidate models within a predefined search space. In practi

ce, a one-shot supernet is trained to serve as an evaluator. A faithful ranking certainly leads to more accurate searching results. However, current methods are prone to making misjudgments. In this paper, we prove that their biased evaluation is due to inherent unfairness in the supernet training. In view of this, we propose two levels of constraints: expectation fairness and strict fairness. Particularly, strict fairness ensures equal optimization opportunities for all choice blocks throughout the training, which neither overestimates nor underestimates their capacity. We demonstrate that this is crucial for improving the confidence of models' ranking. Incorporating the one-shot supernet trained under the proposed fairness constraints with a multi-objective evolutionary search algorithm, we obtain various state-of-the-art models, e.g., FairNAS-A attains 77.5% top-1 validation accuracy on ImageNet.

****

Keep CALM and Improve Visual Feature Attribution

Jae Myung Kim, Junsuk Choe, Zeynep Akata, Seong Joon Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8350-8360

The class activation mapping, or CAM, has been the cornerstone of feature attribution methods for multiple vision tasks. Its simplicity and effectiveness have led to wide applications in the explanation of visual predictions and weakly-supervised localization tasks. However, CAM has its own shortcomings. The computation of attribution maps relies on ad-hoc calibration steps that are not part of the training computational graph, making it difficult for us to understand the real meaning of the attribution values. In this paper, we improve CAM by explicitly incorporating a latent variable encoding the location of the cue for recognition in the formulation, thereby subsuming the attribution map into the training computational graph. The resulting model, class activation latent mapping, or CALM, is trained with the expectation-maximization algorithm. Our experiments show that CALM identifies discriminative attributes for image classifiers more accurately than CAM and other visual attribution baselines. CALM also shows performance improvements over prior arts on the weakly-supervised object localization benchmarks. Our code is available at https://github.com/naver-ai/calm.

****

AdaMML: Adaptive Multi-Modal Learning for Efficient Video Recognition

Rameswar Panda, Chun-Fu (Richard) Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, Rogerio Feris; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7576-7585

Multi-modal learning, which focuses on utilizing various modalities to improve the performance of a model, is widely used in video recognition. While traditional multi-modal learning offers excellent recognition results, its computational expense limits its impact for many real-world applications. In this paper, we propose an adaptive multi-modal learning framework, called AdaMML, that selects on-the-fly the optimal modalities for each segment conditioned on the input for efficient video recognition. Specifically, given a video segment, a multi-modal policy network is used to decide what modalities should be used for processing by the recognition model, with the goal of improving both accuracy and efficiency. We efficiently train the policy network jointly with the recognition model using standard back-propagation. Extensive experiments on four challenging diverse datasets demonstrate that our proposed adaptive approach yields 35%-55% reduction in computation when compared to the traditional baseline that simply uses all the modalities irrespective of the input, while also achieving consistent improvements in accuracy over the state-of-the-art methods. Project page: https://rpand002.github.io/adamml.html.

****

Rethinking 360deg Image Visual Attention Modelling With Unsupervised Learning.

Yasser Abdelaziz Dahou Djilali, Tarun Krishna, Kevin McGuinness, Noel E. O'Connor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15414-15424

Despite the success of self-supervised representation learning on planar data, to date it has not been studied on 360deg images. In this paper, we extend recent advances in contrastive learning to learn latent representations that are suffi

ciently invariant to be highly effective for spherical saliency prediction as a downstream task. We argue that omni-directional images are particularly suited to such an approach due to the geometry of the data domain. To verify this hypothesis, we design an unsupervised framework that effectively maximizes the mutual information between the different views from both the equator and the poles. We show that the decoder is able to learn good quality saliency distributions from the encoder embeddings. Our model compares favorably with fully-supervised learning methods on the Salient360!, VR-EyeTracking and Sitzman datasets. This performance is achieved using an encoder that is trained in a completely unsupervised way and a relatively lightweight supervised decoder (3.8 X fewer parameters in the case of the ResNet50 encoder). We believe that this combination of supervised and unsupervised learning is an important step toward flexible formulations of human visual attention.

****

An End-to-End Transformer Model for 3D Object Detection

Ishan Misra, Rohit Girdhar, Armand Joulin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2906-2917

We propose 3DETR, an end-to-end Transformer based object detection model for 3D point clouds. Compared to existing detection methods that employ a number of 3D-specific inductive biases, 3DETR requires minimal modifications to the vanilla Transformer block. Specifically, we find that a standard Transformer with non-parametric queries and Fourier positional embeddings is competitive with specialized architectures that employ libraries of 3D-specific operators with hand-tuned hyperparameters. Nevertheless, 3DETR is conceptually simple and easy to implement, enabling further improvements by incorporating 3D domain knowledge. Through extensive experiments, we show 3DETR outperforms the well-established and highly optimized VoteNet baselines on the challenging ScanNetV2 dataset by 9.5%. Furthermore, we show 3DETR is applicable to 3D tasks beyond detection, and can serve as a building block for future research.

****

Lipschitz Continuity Guided Knowledge Distillation

Yuzhang Shang, Bin Duan, Ziliang Zong, Liqiang Nie, Yan Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10675-10684

Knowledge distillation has become one of the most important model compression techniques by distilling knowledge from larger teacher networks to smaller student ones. Although great success has been achieved by prior distillation methods via delicately designing various types of knowledge, they overlook the functional properties of neural networks, which makes the process of applying those techniques to new tasks unreliable and non-trivial. To alleviate such problem, in this paper, we initially leverage Lipschitz continuity to better represent the functional characteristic of neural networks and guide the knowledge distillation process. In particular, we propose a novel Lipschitz Continuity Guided Knowledge Distillation framework to faithfully distill knowledge by minimizing the distance between two neural networks' Lipschitz constants, which enables teacher networks to better regularize student networks and improve the corresponding performance. We derive an explainable approximation algorithm with an explicit theoretical derivation to address the NP-hard problem of calculating the Lipschitz constant. Experimental results have shown that our method outperforms other benchmarks over several knowledge distillation tasks (e.g., classification, segmentation and object detection) on CIFAR-100, ImageNet, and PASCAL VOC datasets.

****

Instance Similarity Learning for Unsupervised Feature Representation

Ziwei Wang, Yunsong Wang, Ziyi Wu, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10336-10345

In this paper, we propose an instance similarity learning (ISL) method for unsupervised feature representation. Conventional methods assign close instance pairs in the feature space with high similarity, which usually leads to wrong pairwise relationship for large neighborhoods because the Euclidean distance fails to depict the true semantic similarity on the feature manifold. On the contrary, our

method mines the feature manifold in an unsupervised manner, through which the semantic similarity among instances is learned in order to obtain discriminative representations. Specifically, we employ the Generative Adversarial Networks (GAN) to mine the underlying feature manifold, where the generated features are applied as the proxies to progressively explore the feature manifold so that the semantic similarity among instances is acquired as reliable pseudo supervision. Extensive experiments on image classification demonstrate the superiority of our method compared with the state-of-the-art methods. The code is available at https://github.com/ZiweiWangTHU/ISL.git.

****

Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives

Ben Saunders, Necati Cihan Camgoz, Richard Bowden; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1919-1929

It is common practice to represent spoken languages at their phonetic level. However, for sign languages, this implies breaking motion into its constituent motion primitives. Avatar based Sign Language Production (SLP) has traditionally done just this, building up animation from sequences of hand motions, shapes and facial expressions. However, more recent deep learning based solutions to SLP have tackled the problem using a single network that estimates the full skeletal structure. We propose splitting the SLP task into two distinct jointly-trained sub-tasks. The first translation sub-task translates from spoken language to a latent sign language representation, with gloss supervision. Subsequently, the animation sub-task aims to produce expressive sign language sequences that closely resemble the learnt spatio-temporal representation. Using a progressive transformer for the translation sub-task, we propose a novel Mixture of Motion Primitives (MoMP) architecture for sign language animation. A set of distinct motion primitives are learnt during training, that can be temporally combined at inference to animate continuous sign language sequences. We evaluate on the challenging RWTH-PHOENIX-Weather-2014T(PHOENIX14T) dataset, presenting extensive ablation studies and showing that MoMP outperforms baselines in user evaluations. We achieve state-of-the-art back translation performance with an 11% improvement over competing results. Importantly, and for the first time, we showcase stronger performance for a full translation pipeline going from spoken language to sign, than from gloss to sign.

****

DocFormer: End-to-End Transformer for Document Understanding

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, R. Manmatha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 993-1003

We present DocFormer - a multi-modal transformer based architecture for the task of Visual Document Understanding (VDU). VDU is a challenging problem which aims to understand documents in their varied formats(forms, receipts etc.) and layouts. In addition, DocFormer is pre-trained in an unsupervised fashion using carefully designed tasks which encourage multi-modal interaction. DocFormer uses text, vision and spatial features and combines them using a novel multi-modal self-attention layer. DocFormer also shares learned spatial embeddings across modalities which makes it easy for the model to correlate text to visual tokens and vice versa. DocFormer is evaluated on 4 different datasets each with strong baselines. DocFormer achieves state-of-the-art results on all of them, sometimes beating models 4x its size (in no. of parameters)

****

Spatially-Adaptive Image Restoration Using Distortion-Guided Networks

Kuldeep Purohit, Maitreya Suin, A. N. Rajagopalan, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2309-2319

We present a general learning-based solution for restoring images suffering from spatially-varying degradations. Prior approaches are typically degradation-specific and employ the same processing across different images and different pixels within. However, we hypothesize that such spatially rigid processing is suboptimal for simultaneously restoring the degraded pixels as well as reconstructing t

he clean regions of the image. To overcome this limitation, we propose SPAIR, a network design that harnesses distortion-localization information and dynamically adjusts computation to difficult regions in the image. SPAIR comprises of two components, (1) a localization network that identifies degraded pixels, and (2) a restoration network that exploits knowledge from the localization network in filter and feature domain to selectively and adaptively restore degraded pixels. Our key idea is to exploit the non-uniformity of heavy degradations in spatial-domain and suitably embed this knowledge within distortion-guided modules performing sparse normalization, feature extraction and attention. Our architecture is agnostic to physical formation model and generalizes across several types of spatially-varying degradations. We demonstrate the efficacy of SPAIR individually on four restoration tasks- removal of rain-streaks, raindrops, shadows and motion blur. Extensive qualitative and quantitative comparisons with prior art on 11 benchmark datasets demonstrate that our degradation-agnostic network design offers significant performance gains over state-of-the-art degradation-specific architectures. Code available at https://github.com/human-analysis/spatially-adaptive-image-restoration.

****

Exploiting Sample Correlation for Crowd Counting With Multi-Expert Network

Xinyan Liu, Guorong Li, Zhenjun Han, Weigang Zhang, Yifan Yang, Qingming Huang, Nicu Sebe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3215-3224

Crowd counting is a difficult task because of the diversity of scenes. Most of the existing crowd counting methods adopt complex structures with massive backbones to enhance the generalization ability. Unfortunately, the performance of existing methods on large-scale data sets is not satisfactory. In order to handle various scenarios with less complex network, we explored how to efficiently use the multi-expert model for crowd counting tasks. We mainly focus on how to train more efficient expert networks and how to choose the most suitable expert. Specifically, we propose a task-driven similarity metric based on sample's mutual enhancement, referred as co-fine-tune similarity, which can find a more efficient subset of data for training the expert network. Similar samples are considered as a cluster which is used to obtain parameters of an expert. Besides, to make better use of the proposed method, we design a simple network called FPN with Deconvolution Counting Network, which is a more suitable base model for the multi-expert counting network. Experimental results show that multiple experts FDC (MFDC) achieves the best performance on four public data sets, including the large scale NWPU-Crowd data set. Furthermore, the MFDC trained on an extensive dense crowd data set can generalize well on the other data sets without extra training or fine-tuning.

****

Unlimited Neighborhood Interaction for Heterogeneous Trajectory Prediction

Fang Zheng, Le Wang, Sanping Zhou, Wei Tang, Zhenxing Niu, Nanning Zheng, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13168-13177

Understanding complex social interactions among agents is a key challenge for trajectory prediction. Most existing methods consider the interactions between pairwise traffic agents or in a local area, while the nature of interactions is unlimited, involving an uncertain number of agents and non-local areas simultaneously. Besides, they treat heterogeneous traffic agents the same, namely those among agents of different categories, while neglecting people's diverse reaction patterns toward traffic agents in different categories. To address these problems, we propose a simple yet effective Unlimited Neighborhood Interaction Network (UNIN), which predicts trajectories of heterogeneous agents in multiple categories. Specifically, the proposed unlimited neighborhood interaction module generates the fused-features of all agents involved in an interaction simultaneously, which is adaptive to any number of agents and any range of interaction area. Meanwhile, a hierarchical graph attention module is proposed to obtain category-to-category interaction and agent-to-agent interaction. Finally, parameters of a Gaussian Mixture Model are estimated for generating the future trajectories. Extensive

experimental results on benchmark datasets demonstrate a significant performanc
e improvement of our method over the state-of-the-art methods.
****

Multi-Scale Matching Networks for Semantic Correspondence
Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, Yizhou Yu; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
021, pp. 3354-3364
Deep features have been proven powerful in building accurate dense semantic corr
espondences in various previous works. However, the multi-scale and pyramidal hi
erarchy of convolutional neural networks has not been well studied to learn disc
riminative pixel-level features for semantic correspondence. In this paper, we p
ropose a multiscale matching network that is sensitive to tiny semantic differen
ces between neighboring pixels. We follow the coarse-to-fine matching strategy,
and build a top-down feature and matching enhancement scheme that is coupled wit
h the multi-scale hierarchy of deep convolutional neural networks. During featur
e enhancement, intra-scale enhancement fuses same-resolution feature maps from m
ultiple layers together via local self-attention, and cross-scale enhancement ha
llucinates higher resolution feature maps along the top-down hierarchy. Besides,
 we learn complementary matching details at different scales, and thus the overa
ll matching score is refined by features at different semantic levels gradually.
 Our multi-scale matching network can be trained end-to-end easily with few addi
tional learnable parameters. Experimental results demonstrate the proposed metho
d achieves state-of-the-art performance on three popular benchmarks with high co
mputational efficiency.
****

LatentCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpr
etable Directions
O■uz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, Pinar Yanardag; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14263-
14272
Recent research has shown that it is possible to find interpretable directions i
n the latent spaces of pre-trained Generative Adversarial Networks (GANs). These
 directions enable controllable image generation and support a wide range of sem
antic editing operations, such as zoom or rotation. The discovery of such direct
ions is often done in a supervised or semi-supervised manner and requires manual
 annotations which limits their use in practice. In comparison, unsupervised dis
covery allows finding subtle directions that are difficult to detect a priori. I
n this work, we propose a contrastive learning-based approach to discover semant
ic directions in the latent space of pre-trained GANs in a self-supervised manne
r. Our approach finds semantically meaningful dimensions compatible with state-o
f-the-art methods.
****

Few-Shot Visual Relationship Co-Localization
Revant Teotia, Vaibhav Mishra, Mayank Maheshwari, Anand Mishra; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16342-
16351
In this paper, given a small bag of images, each containing a common but latent
predicate, we are interested in localizing visual subject-object pairs connected
 via the common predicate in each of the images. We refer to this novel problem
as visual relationship co-localization or VRC as an abbreviation. VRC is a chall
enging task, even more so than the well-studied object co-localization task. Thi
s becomes further challenging when using just a few images, the model has to lea
rn to co-localize visual subject-object pairs connected via unseen predicates. T
o solve VRC, we propose an optimization framework to select a common visual rela
tionship in each image of the bag. The goal of the optimization framework is to
find the optimal solution by learning visual relationship similarity across imag
es in a few-shot setting. To obtain robust visual relationship representation, w
e utilize a simple yet effective technique that learns relationship embedding as
 a translation vector from visual subject to visual object in a shared space. Fu
rther, to learn visual relationship similarity, we utilize a proven meta-learnin

g technique commonly used for few-shot classification tasks. Finally, to tackle the combinatorial complexity challenge arising from an exponential number of feasible solutions, we use a greedy approximation inference algorithm that selects approximately the best solution. We extensively evaluate our proposed framework on variations of bag sizes obtained from two challenging public datasets, namely VrR-VG and VG-150, and achieve impressive visual co-localization performance.
****

RFNet: Recurrent Forward Network for Dense Point Cloud Completion
Tianxin Huang, Hao Zou, Jinhao Cui, Xuemeng Yang, Mengmeng Wang, Xiangrui Zhao, Jiangning Zhang, Yi Yuan, Yifan Xu, Yong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12508-12517
Point cloud completion is an interesting and challenging task in 3D vision, aiming to recover complete shapes from sparse and incomplete point clouds. Existing learning-based methods often require vast computation cost to achieve excellent performance, which limits their practical applications. In this paper, we propose a novel Recurrent Forward Network (RFNet), which is composed of three modules: Recurrent Feature Extraction (RFE), Forward Dense Completion (FDC) and Raw Shape Protection (RSP). The RFE extracts multiple global features from the incomplete point clouds for different recurrent levels, and the FDC generates point clouds in a coarse-to-fine pipeline. The RSP introduces details from the original incomplete models to refine the completion results. Besides, we propose a Sampling Chamfer Distance to better capture the shapes of models and a new Balanced Expansion Constraint to restrict the expansion distances from coarse to fine. According to the experiments on ShapeNet and KITTI, our network can achieve the state-of-the-art with lower memory cost and faster convergence.
****

Towards Better Explanations of Class Activation Mapping
Hyungsik Jung, Youngrock Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1336-1344
Increasing demands for understanding the internal behavior of convolutional neural networks (CNNs) have led to remarkable improvements in explanation methods. Particularly, several class activation mapping (CAM) based methods, which generate visual explanation maps by a linear combination of activation maps from CNNs, have been proposed. However, the majority of the methods lack a clear theoretical basis on how they assign the coefficients of the linear combination. In this paper, we revisit the intrinsic linearity of CAM with respect to the activation maps; we construct an explanation model of CNN as a linear function of binary variables that denote the existence of the corresponding activation maps. With this approach, the explanation model can be determined by additive feature attribution methods in an analytic manner. We then demonstrate the adequacy of SHAP values, which is a unique solution for the explanation model with a set of desirable properties, as the coefficients of CAM. Since the exact SHAP values are unattainable, we introduce an efficient approximation method, LIFT-CAM, based on DeepLIFT. Our proposed LIFT-CAM can estimate the SHAP values of the activation maps with high speed and accuracy. Furthermore, it greatly outperforms other previous CAM-based methods in both qualitative and quantitative aspects.
****

Domain Adaptive Video Segmentation via Temporal Consistency Regularization
Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8053-8064
Video semantic segmentation is an essential task for the analysis and understanding of videos. Recent efforts largely focus on supervised video segmentation by learning from fully annotated data, but the learnt models often experience clear performance drop while applied to videos of a different domain. This paper presents DA-VSN, a domain adaptive video segmentation network that addresses domain gaps in videos by temporal consistency regularization (TCR) for consecutive frames of target-domain videos. DA-VSN consists of two novel and complementary designs. The first is cross-domain TCR that guides the prediction of target frames to have similar temporal consistency as that of source frames (learnt from annotated source data) via adversarial learning. The second is intra-domain TCR that gu

ides unconfident predictions of target frames to have similar temporal consisten cy as confident predictions of target frames. Extensive experiments demonstrate the superiority of our proposed domain adaptive video segmentation network which outperforms multiple baselines consistently by large margins.
****

PR-Net: Preference Reasoning for Personalized Video Highlight Detection
Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglun Chen, Pai Peng, Xing Sun, Wenpin g Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7980-7989
Personalized video highlight detection aims to shorten a long video to interesti ng moments according to a user's preference, which has recently raised the commu nity's attention. Current methods regard the user's history as holistic informat ion to predict the user's preference but negating the inherent diversity of the user's interests, resulting in vague preference representation. In this paper, w e propose a simple yet efficient preference reasoning framework (PR-Net) to expl icitly take the diverse interests into account for frame-level highlight predict ion. Specifically, distinct user-specific preferences for each input query frame are produced, presented as the similarity weighted sum of history highlights to the corresponding query frame. Next, distinct comprehensive preferences are for med by the user-specific preferences and a learnable generic preference for more overall highlight measurement. Lastly, the degree of highlight and non-highligh t for each query frame is calculated as semantic similarity to its comprehensive and non-highlight preferences, respectively. Besides, to alleviate the ambiguit y due to the incomplete annotation, a new bi-directional contrastive loss is pro posed to ensure a compact and differentiable metric space. In this way, our meth od significantly outperforms state-of-the-art methods with a relative improvemen t of 12% in mean accuracy precision.
****

PoinTr: Diverse Point Cloud Completion With Geometry-Aware Transformers
Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1249 8-12507
Point clouds captured in real-world applications are often incomplete due to the limited sensor resolution, single viewpoint, and occlusion. Therefore, recoveri ng the complete point clouds from partial ones becomes an indispensable task in many practical applications. In this paper, we present a new method that reformu lates point cloud completion as a set-to-set translation problem and design a ne w model, called PoinTr that adopts a transformer encoder-decoder architecture fo r point cloud completion. By representing the point cloud as a set of unordered groups of points with position embeddings, we convert the point cloud to a seque nce of point proxies and employ the transformers for point cloud generation. To facilitate transformers to better leverage the inductive bias about 3D geometric structures of point clouds, we further devise a geometry-aware block that model s the local geometric relationships explicitly. The migration of transformers en ables our model to better learn structural knowledge and preserve detailed infor mation for point cloud completion. Furthermore, we propose two more challenging benchmarks with more diverse incomplete point clouds that can better reflect the real-world scenarios to promote future research. Experimental results show that our method outperforms state-of-the-art methods by a large margin on both the n ew benchmarks and the existing ones.
****

Learn-To-Race: A Multimodal Control Environment for Autonomous Racing
James Herman, Jonathan Francis, Siddha Ganju, Bingqing Chen, Anirudh Koul, Abhin av Gupta, Alexey Skabelkin, Ivan Zhukov, Max Kumskoy, Eric Nyberg; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 979 3-9802
Existing research on autonomous driving primarily focuses on urban driving, whic h is insufficient for characterising the complex driving behaviour underlying hi gh-speed racing. At the same time, existing racing simulation frameworks struggl e in capturing realism, with respect to visual rendering, vehicular dynamics, an

d task objectives, inhibiting the transfer of learning agents to real-world cont
exts. We introduce a new environment, where agents Learn-to-Race (L2R) in simula
ted competition-style racing, using multimodal information|from virtual cameras
to a comprehensive array of inertial measurement sensors. Our environment, which
 includes a simulator and an interfacing training framework, accurately models v
ehicle dynamics and racing conditions. In this paper, we release the Arrival sim
ulator for autonomous racing. Next, we propose the L2R task with challenging met
rics, inspired by learning-to-drive challenges, Formula-style racing, and multim
odal trajectory prediction for autonomous driving. Additionally, we provide the
L2R framework suite, facilitating simulated racing on high-precision models of r
eal-world tracks. Finally, we provide an official L2R task dataset of expert dem
onstrations, as well as a series of baseline experiments and reference implement
ations. We make all code available: https://github.com/learn-to-race/l2r.
****

SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Reco
gnition
Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, Houqiang Li; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1108
7-11096
Hand gesture serves as a critical role in sign language. Current deep-learning-b
ased sign language recognition (SLR) methods may suffer insufficient interpretab
ility and overfitting due to limited sign data sources. In this paper, we introd
uce the first self-supervised pre-trainable SignBERT with incorporated hand prio
r for SLR. SignBERT views the hand pose as a visual token, which is derived from
 an off-the-shelf pose extractor. The visual tokens are then embedded with gestu
re state, temporal and hand chirality information. To take full advantage of ava
ilable sign data sources, SignBERT first performs self-supervised pre-training b
y masking and reconstructing visual tokens. Jointly with several mask modeling s
trategies, we attempt to incorporate hand prior in a model-aware method to bette
r model hierarchical context over the hand sequence. Then with the prediction he
ad added, SignBERT is fine-tuned to perform the downstream SLR task. To validate
 the effectiveness of our method on SLR, we perform extensive experiments on fou
r public benchmark datasets, i.e., NMFs-CSL, SLR500, MSASL and WLASL. Experiment
 results demonstrate the effectiveness of both self-supervised learning and impo
rted hand prior. Furthermore, we achieve state-of-the-art performance on all ben
chmarks with a notable gain.
****

Improving Low-Precision Network Quantization via Bin Regularization
Tiantian Han, Dong Li, Ji Liu, Lu Tian, Yi Shan; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 5261-5270
Model quantization is an important mechanism for energy-efficient deployment of
deep neural networks on resource-constrained devices by reducing the bit precisi
on of weights and activations. However, it remains challenging to maintain high
accuracy as bit precision decreases, especially for low-precision networks (e.g.
, 2-bit MobileNetV2). Existing methods have explored to address this problem by
minimizing the quantization error or mimicking the data distribution of full-pre
cision networks. In this work, we propose a novel weight regularization algorith
m for improving low-precision network quantization. Instead of constraining the
overall data distribution, we separably optimize all elements in each quantizati
on bin to be as close to the target quantized value as possible. Such bin regula
rization (BR) mechanism encourages the weight distribution of each quantization
bin to be sharp and approximate to a Dirac delta distribution ideally. Experimen
ts demonstrate that our method achieves consistent improvements over the state-o
f-the-art quantization-aware training methods for different low-precision networ
ks. Particularly, our bin regularization improves LSQ for 2-bit MobileNetV2 and
MobileNetV3-Small by 3.9% and 4.9% top-1 accuracy on ImageNet, respectively.
****

Probabilistic Modeling for Human Mesh Recovery
Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, Kostas Daniilidis; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,

pp. 11605-11614
This paper focuses on the problem of 3D human reconstruction from 2D evidence. A lthough this is an inherently ambiguous problem, the majority of recent works av oid the uncertainty modeling and typically regress a single estimate for a given input. In contrast to that, in this work, we propose to embrace the reconstruct ion ambiguity and we recast the problem as learning a mapping from the input to a distribution of plausible 3D poses. Our approach is based on the normalizing f lows model and offers a series of advantages. For conventional applications, whe re a single 3D estimate is required, our formulation allows for efficient mode c omputation. Using the mode leads to performance that is comparable with the stat e of the art among deterministic unimodal regression models. Simultaneously, sin ce we have access to the likelihood of each sample, we demonstrate that our mode l is useful in a series of downstream tasks, where we leverage the probabilistic nature of the prediction as a tool for more accurate estimation. These tasks in clude reconstruction from multiple uncalibrated views, as well as human model fi tting, where our model acts as a powerful image-based prior for mesh recovery. O ur results validate the importance of probabilistic modeling, and indicate state -of-the-art performance across a variety of settings. Code and models are availa ble at: https://www.seas.upenn.edu/ nkolot/projects/prohmr.
****

Distilling Virtual Examples for Long-Tailed Recognition
Yin-Yin He, Jianxin Wu, Xiu-Shen Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 235-244
We tackle the long-tailed visual recognition problem from the knowledge distilla tion perspective by proposing a Distill the Virtual Examples (DiVE) method. Spec ifically, by treating the predictions of a teacher model as virtual exam- ples, we prove that distilling from these virtual examples is equivalent to label dist ribution learning under certain con- straints. We show that when the virtual exa mple distribu- tion becomes flatter than the original input distribution, the un der-represented tail classes will receive significant im- provements, which is c rucial in long-tailed recognition. The proposed DiVE method can explicitly tune the virtual exam- ple distribution to become flat. Extensive experiments on thre e benchmark datasets, including the large-scale iNat- uralist ones, justify that the proposed DiVE method can significantly outperform state-of-the-art methods. Further- more, additional analyses and experiments verify the virtual example i nterpretation, and demonstrate the effectiveness of tailored designs in DiVE for long-tailed problems.
****

Understanding and Mitigating Annotation Bias in Facial Expression Recognition
Yunliang Chen, Jungseock Joo; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2021, pp. 14980-14991
The performance of a computer vision model depends on the size and quality of it s training data. Recent studies have unveiled previously-unknown composition bia ses in common image datasets which then lead to skewed model outputs, and have p roposed methods to mitigate these biases. However, most existing works assume th at human-generated annotations can be considered gold-standard and unbiased. In this paper, we reveal that this assumption can be problematic, and that special care should be taken to prevent models from learning such annotation biases. We focus on facial expression recognition and compare the label biases between lab-controlled and in-the-wild datasets. We demonstrate that many expression dataset s contain significant annotation biases between genders, especially when it come s to the happy and angry expressions, and that traditional methods cannot fully mitigate such biases in trained models. To remove expression annotation bias, we propose an AU-Calibrated Facial Expression Recognition (AUC-FER) framework that utilizes facial action units (AUs) and incorporates the triplet loss into the o bjective function. Experimental results suggest that the proposed method is more effective in removing expression annotation bias than existing techniques.
****

Large Scale Multi-Illuminant (LSMI) Dataset for Developing White Balance Algorit hm Under Mixed Illumination

Dongyoung Kim, Jinwoo Kim, Seonghyeon Nam, Dongwoo Lee, Yeonkyung Lee, Nahyup Kang, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2410-2419
We introduce a Large Scale Multi-Illuminant (LSMI) Dataset that contains 7,486 images, captured with three different cameras on more than 2,700 scenes with two or three illuminants. For each image in the dataset, the new dataset provides not only the pixel-wise ground truth illumination but also the chromaticity of each illuminant in the scene and the mixture ratio of illuminants per pixel. Images in our dataset are mostly captured with illuminants existing in the scene, and the ground truth illumination is computed by taking the difference between the images with different illumination combination. Therefore, our dataset captures natural composition in the real-world setting with wide field-of-view, providing more extensive dataset compared to existing datasets for multi-illumination white balance. As conventional single illuminant white balance algorithms cannot be directly applied, we also apply per-pixel DNN-based white balance algorithm and show its effectiveness against using patch-wise white balancing. We validate the benefits of our dataset through extensive analysis including a user-study, and expect the dataset to make meaningful contribution for future work in white balancing.

****

Reality Transform Adversarial Generators for Image Splicing Forgery Detection and Localization
Xiuli Bi, Zhipeng Zhang, Bin Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14294-14303
When many forged images become more and more realistic with the help of image editing tools and deep learning techniques, authenticators need to improve their ability to verify these forged images. The process of generating and detecting forged images is thus similar to the principle of Generative Adversarial Networks (GANs). Creating realistic forged images requires a retouching process to suppress tampering artifacts and keep structural information. We view this retouching process as image style transfer and then proposed the fake-to-realistic transformation generator GT. For detecting the tampered regions, a forgery localization generator GM is proposed based on a multi-decoder-single-task strategy. By adversarial training two generators, the proposed alpha-learnable whitening and coloring transformation (alpha-learnable WCT) block in GT automatically suppresses the tampering artifacts in the forged images. Meanwhile, the detection and localization abilities of GM will be improved by learning the forged images retouched by GT. The experimental results demonstrate that the proposed two generators in GAN can simulate confrontation between fakers and authenticators well. The localization generator GM outperforms the state-of-the-art methods in splicing forgery detection and localization on four public datasets.

****

Learning To Regress Bodies From Images Using Differentiable Semantic Rendering
Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11250-11259
Learning to regress 3D human body shape and pose (e.g. SMPL parameters) from monocular images typically exploits losses on 2D keypoints, silhouettes, and/or part-segmentation when 3D training data is not available. Such losses, however, are limited because 2D keypoints do not supervise body shape and segmentations of people in clothing do not match projected minimally-clothed SMPL shapes. To exploit richer image information about clothed people, we introduce higher-level semantic information about clothing to penalize clothed and non-clothed regions of the image differently. To do so, we train a body regressor using a novel "Differentiable Semantic Rendering - DSR" loss. For Minimally-Clothed regions, we define the DSR-MC loss, which encourages a tight match between a rendered SMPL body and the minimally-clothed regions of the image. For clothed regions, we define the DSR-C loss to encourage the rendered SMPL body to be inside the clothing mask. To ensure end-to-end differentiable training, we learn a semantic clothing prior for SMPL vertices from thousands of clothed human scans. We perform extensive q

ualitative and quantitative experiments to evaluate the role of clothing semanti cs on the accuracy of 3D human pose and shape estimation. We outperform all prev ious state-of-the-art methods on 3DPW and Human3.6M and obtain on par results on MPI-INF-3DHP. Code and trained models will be available for research at https:/ /dsr.is.tue.mpg.de/
****

Bifold and Semantic Reasoning for Pedestrian Behavior Prediction
Amir Rasouli, Mohsen Rohani, Jun Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15600-15610
Pedestrian behavior prediction is one of the major challenges for intelligent dr iving systems. Pedestrians often exhibit complex behaviors influenced by various contextual elements. To address this problem, we propose BiPed, a multitask lea rning framework that simultaneously predicts trajectories and actions of pedestr ians by relying on multimodal data. Our method benefits from 1) a bifold encodin g approach where different data modalities are processed independently allowing them to develop their own representations, and jointly to produce a representati on for all modalities using shared parameters; 2) a novel interaction modeling t echnique that relies on categorical semantic parsing of the scenes to capture in teractions between target pedestrians and their surroundings; and 3) a bifold pr ediction mechanism that uses both independent and shared decoding of multimodal representations. Using public pedestrian behavior benchmark datasets for driving , PIE and JAAD, we highlight the benefits of the proposed method for behavior pr ediction and show that our model achieves state-of-the-art performance and impro ves trajectory and action prediction by up to 22% and 9% respectively. We furthe r investigate the contributions of the proposed reasoning techniques via extensi ve ablation studies.
****

Learning Target Candidate Association To Keep Track of What Not To Track
Christoph Mayer, Martin Danelljan, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13 444-13454
The presence of objects that are confusingly similar to the tracked target, pose s a fundamental challenge in appearance-based visual tracking. Such distractor o bjects are easily misclassified as the target itself, leading to eventual tracki ng failure. While most methods strive to suppress distractors through more power ful appearance models, we take an alternative approach. We propose to keep track of distractor objects in order to continue tracking the target. To this end, we introduce a learned association network, allowing us to propagate the identitie s of all target candidates from frame-to-frame. To tackle the problem of lacking ground-truth correspondences between distractor objects in visual tracking, we propose a training strategy that combines partial annotations with self-supervis ion. We conduct comprehensive experimental validation and analysis of our approa ch on several challenging datasets. Our tracker sets a new state-of-the-art on s ix benchmarks, achieving an AUC score of 67.1% on LaSOT and a +5.8% absolute gai n on the OxUvA long-term dataset.
****

VolumeFusion: Deep Depth Fusion for 3D Scene Reconstruction
Jaesung Choe, Sunghoon Im, Francois Rameau, Minjun Kang, In So Kweon; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16086-16095
To reconstruct a 3D scene from a set of calibrated views, traditional multi-view stereo techniques rely on two distinct stages: local depth maps computation and global depth maps fusion. Recent studies concentrate on deep neural architectur es for depth estimation by using conventional depth fusion method or direct 3D r econstruction network by regressing Truncated Signed Distance Function (TSDF). I n this paper, we advocate that replicating the traditional two stages framework with deep neural networks improves both the interpretability and the accuracy of the results. As mentioned, our network operates in two steps: 1) the local comp utation of the local depth maps with a deep MVS technique, and, 2) the depth map s and images' features fusion to build a single TSDF volume. In order to improve

the matching performance between images acquired from very different viewpoints (e.g., large-baseline and rotations), we introduce a rotation-invariant 3D convolution kernel called PosedConv. The effectiveness of the proposed architecture is underlined via a large series of experiments conducted on the ScanNet dataset where our approach compares favorably against both traditional and deep learning techniques.

****

Black-Box Detection of Backdoor Attacks With Limited Information and Data

Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, Jun Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16482-16491

Although deep neural networks (DNNs) have made rapid progress in recent years, they are vulnerable in adversarial environments. A malicious backdoor could be embedded in a model by poisoning the training dataset, whose intention is to make the infected model give wrong predictions during inference when the specific trigger appears. To mitigate the potential threats of backdoor attacks, various backdoor detection and defense methods have been proposed. However, the existing techniques usually require the poisoned training data or access to the white-box model, which is commonly unavailable in practice. In this paper, we propose a black-box backdoor detection (B3D) method to identify backdoor attacks with only query access to the model. We introduce a gradient-free optimization algorithm to reverse-engineer the potential trigger for each class, which helps to reveal the existence of backdoor attacks. In addition to backdoor detection, we also propose a simple strategy for reliable predictions using the identified backdoored models. Extensive experiments on hundreds of DNN models trained on several datasets corroborate the effectiveness of our method under the black-box setting against various backdoor attacks.

****

A Robust Loss for Point Cloud Registration

Zhi Deng, Yuxin Yao, Bailin Deng, Juyong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6138-6147

The performance of surface registration relies heavily on the metric used for the alignment error between the source and target shapes. Traditionally, such a metric is based on the point-to-point or point-to-plane distance from the points on the source surface to their closest points on the target surface, which is susceptible to failure due to instability of the closest-point correspondence. In this paper, we propose a novel metric based on the intersection points between the two shapes and a random straight line, which does not assume a specific correspondence. We verify the effectiveness of this metric by extensive experiments, including its direct optimization for a single registration problem as well as unsupervised learning for a set of registration problems. The results demonstrate that the algorithms utilizing our proposed metric outperforms the state-of-the-art optimization-based and unsupervised learning-based methods.

****

Semantic Concentration for Domain Adaptation

Shuang Li, Mixue Xie, Fangrui Lv, Chi Harold Liu, Jian Liang, Chen Qin, Wei Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9102-9111

Domain adaptation (DA) paves the way for label annotation and dataset bias issues by the knowledge transfer from a label-rich source domain to a related but unlabeled target domain. A mainstream of DA methods is to align the feature distributions of the two domains. However, the majority of them focus on the entire image features where irrelevant semantic information, e.g., the messy background, is inevitably embedded. Enforcing feature alignments in such case will negatively influence the correct matching of objects and consequently lead to the semantically negative transfer due to the confusion of irrelevant semantics. To tackle this issue, we propose Semantic Concentration for Domain Adaptation (SCDA), which encourages the model to concentrate on the most principal features via the pair-wise adversarial alignment of prediction distributions. Specifically, we train the classifier to class-wisely maximize the prediction distribution divergence o

f each sample pair, which enables the model to find the region with large differ
ences among the same class of samples. Meanwhile, the feature extractor attempts
 to minimize that discrepancy, which suppresses the features of dissimilar regio
ns among the same class of samples and accentuates the features of principal par
ts. As a general method, SCDA can be easily integrated into various DA methods a
s a regularizer to further boost their performance. Extensive experiments on the
 cross-domain benchmarks show the efficacy of SCDA.
****

PICCOLO: Point Cloud-Centric Omnidirectional Localization
Junho Kim, Changwoon Choi, Hojun Jang, Young Min Kim; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2021, pp. 3313-3323
We present PICCOLO, a simple and efficient algorithm for omnidirectional localiz
ation. Given a colored point cloud and a 360 panorama image of a scene, our obje
ctive is to recover the camera pose at which the panorama image is taken. Our pi
peline works in an off-the-shelf manner with a single image given as a query and
 does not require any training of neural networks or collecting ground-truth pos
es of images. Instead, we match each point cloud color to the holistic view of t
he panorama image with gradient-descent optimization to find the camera pose. Ou
r loss function, called sampling loss, is point cloud-centric, evaluated at the
projected location of every point in the point cloud. In contrast, conventional
photometric loss is image-centric, comparing colors at each pixel location. With
 a simple change in the compared entities, sampling loss effectively overcomes t
he severe visual distortion of omnidirectional images, and enjoys the global con
text of the 360 view to handle challenging scenarios for visual localization. PI
CCOLO outperforms existing omnidirectional localization algorithms in both accur
acy and stability when evaluated in various environments.
****

Distributional Robustness Loss for Long-Tail Learning
Dvir Samuel, Gal Chechik; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2021, pp. 9495-9504
Real-world data is often unbalanced and long-tailed, but deep models struggle to
 recognize rare classes in the presence of frequent classes. To address unbalanc
ed data, most studies try balancing the data, the loss, or the classifier to red
uce classification bias towards head classes. Far less attention has been given
to the latent representations learned with unbalanced data. We show that the fea
ture extractor part of deep networks suffers greatly from this bias. We propose
a new loss based on robustness theory, which encourages the model to learn high-
quality representations for both head and tail classes. While the general form o
f the robustness loss may be hard to compute, we further derive an easy-to-compu
te upper bound that can be minimized efficiently. This procedure reduces represe
ntation bias towards head classes in the feature space and achieves new SOTA res
ults on CIFAR100-LT, ImageNet-LT, and iNaturalist long-tail benchmarks. We find
that training with robustness increases recognition accuracy of tail classes whi
le largely maintaining the accuracy of head classes. The new robustness loss can
 be combined with various classifier balancing techniques and can be applied to
representations at several layers of the deep model.
****

NGC: A Unified Framework for Learning With Open-World Noisy Data
Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, Yu-Feng Li; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
1, pp. 62-71
The existence of noisy data is prevalent in both the training and testing phases
 of machine learning systems, which inevitably leads to the degradation of model
 performance. There have been plenty of works concentrated on learning with in-d
istribution (IND) noisy labels in the last decade, i.e., some training samples a
re assigned incorrect labels that do not correspond to their true classes. Nonet
heless, in real application scenarios, it is necessary to consider the influence
 of out-of-distribution (OOD) samples, i.e., samples that do not belong to any k
nown classes, which has not been sufficiently explored yet. To remedy this, we s
tudy a new problem setup, namely Learning with Open-world Noisy Data (LOND). The

goal of LOND is to simultaneously learn a classifier and an OOD detector from datasets with mixed IND and OOD noise. In this paper, we propose a new graph-based framework, namely Noisy Graph Cleaning (NGC), which collects clean samples by leveraging geometric structure of data and model predictive confidence. Without any additional training effort, NGC can detect and reject the OOD samples based on the learned class prototypes directly in testing phase. We conduct experiments on multiple benchmarks with different types of noise and the results demonstrate the superior performance of our method against state of the arts.
****

Superpoint Network for Point Cloud Oversegmentation
Le Hui, Jia Yuan, Mingmei Cheng, Jin Xie, Xiaoya Zhang, Jian Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5510-5519
Superpoints are formed by grouping similar points with local geometric structures, which can effectively reduce the number of primitives of point clouds for subsequent point cloud processing. Existing superpoint methods mainly focus on employing clustering or graph partition to generate superpoints with handcrafted or learned features. Nonetheless, these methods cannot learn superpoints of point clouds with an end-to-end network. In this paper, we develop a new deep iterative clustering network to directly generate superpoints from irregular 3D point clouds in an end-to-end manner. Specifically, in our clustering network, we first jointly learn a soft point-superpoint association map from the coordinate and feature spaces of point clouds, where each point is assigned to the superpoint with a learned weight. Furthermore, we then iteratively update the association map and superpoint centers so that we can more accurately group the points into the corresponding superpoints with locally similar geometric structures. Finally, by predicting the pseudo labels of the superpoint centers, we formulate a label consistency loss on the points and superpoint centers to train the network. Extensive experiments on various datasets indicate that our method not only achieves the state-of-the-art on superpoint generation but also improves the performance of point cloud semantic segmentation. Code is available at https://github.com/fpthink/SPNet.
****

Exploring Simple 3D Multi-Object Tracking for Autonomous Driving
Chenxu Luo, Xiaodong Yang, Alan Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10488-10497
3D multi-object tracking in LiDAR point clouds is a key ingredient for self-driving vehicles. Existing methods are predominantly based on the tracking-by-detection pipeline and inevitably require a heuristic matching step for the detection association. In this paper, we present SimTrack to simplify the hand-crafted tracking paradigm by proposing an end-to-end trainable model for joint detection and tracking from raw point clouds. Our key design is to predict the first-appear location of each object in a given snippet to get the tracking identity and then update the location based on motion estimation. In the inference, the heuristic matching step can be completely waived by a simple read-off operation. SimTrack integrates the tracked object association, newborn object detection, and dead track killing in a single unified model. We conduct extensive evaluations on two large-scale datasets: nuScenes and Waymo Open Dataset. Experimental results reveal that our simple approach compares favorably with the state-of-the-art methods while ruling out the heuristic matching rules.
****

Looking Here or There? Gaze Following in 360-Degree Images
Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, Guodong Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3742-3751
Gaze following, i.e., detecting the gaze target of a human subject, in 2D images has become an active topic in computer vision. However, it usually suffers from the out of frame issue due to the limited field-of-view (FoV) of 2D images. In this paper, we introduce a novel task, gaze following in 360-degree images which provide an omnidirectional FoV and can alleviate the out of frame issue. We col

lect the first dataset, "GazeFollow360", for this task, containing around 10,000 360-degree images with complex gaze behaviors under various scenes. Existing 2D gaze following methods suffer from performance degradation in 360-degree images since they may use the assumption that a gaze target is in the 2D gaze sight line. However, this assumption is no longer true for long-distance gaze behaviors in 360-degree images, due to the distortion brought by sphere-to-plane projection. To address this challenge, we propose a 3D sight line guided dual-pathway framework, to detect the gaze target within a local region (here) and from a distant region (there), parallelly. Specifically, the local region is obtained as a 2D cone-shaped field along the 2D projection of the sight line starting at the human subject's head position, and the distant region is obtained by searching along the sight line in 3D sphere space. Finally, the location of the gaze target is determined by fusing the estimations from both the local region and the distant region. Experimental results show that our method achieves significant improvements over previous 2D gaze following methods on our GazeFollow360 dataset.
****

LoOp: Looking for Optimal Hard Negative Embeddings for Deep Metric Learning
Bhavya Vasudeva, Puneesh Deora, Saumik Bhattacharya, Umapada Pal, Sukalpa Chanda ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2021, pp. 10634-10643
Deep metric learning has been effectively used to learn distance metrics for different visual tasks like image retrieval, clustering, etc. In order to aid the training process, existing methods either use a hard mining strategy to extract the most informative samples or seek to generate hard synthetics using an additional network. Such approaches face different challenges and can lead to biased embeddings in the former case, and (i) harder optimization (ii) slower training speed (iii) higher model complexity in the latter case. In order to overcome these challenges, we propose a novel approach that looks for optimal hard negatives (LoOp) in the embedding space, taking full advantage of each tuple by calculating the minimum distance between a pair of positives and a pair of negatives. Unlike mining-based methods, our approach considers the entire space between pairs of embeddings to calculate the optimal hard negatives. Extensive experiments combining our approach and representative metric learning losses reveal a significant boost in performance on three benchmark datasets.
****

How To Train Neural Networks for Flare Removal
Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan , Jonathan T. Barron; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2239-2247
When a camera is pointed at a strong light source, the resulting photograph may contain lens flare artifacts. Flares appear in a wide variety of patterns (halos , streaks, color bleeding, haze, etc.) and this diversity in appearance makes flare removal challenging. Existing analytical solutions make strong assumptions about the artifact's geometry or brightness, and therefore only work well on a small subset of flares. Machine learning techniques have shown success in removing other types of artifacts, like reflections, but have not been widely applied to flare removal due to the lack of training data. To solve this problem, we explicitly model the optical causes of flare either empirically or using wave optics, and generate semi-synthetic pairs of flare-corrupted and clean images. This enables us to train neural networks to remove lens flare for the first time. Experiments show our data synthesis approach is critical for accurate flare removal, and that models trained with our technique generalize well to real lens flares across different scenes, lighting conditions, and cameras.
****

Motion Basis Learning for Unsupervised Deep Homography Estimation With Subspace Projection
Nianjin Ye, Chuan Wang, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13117-13125
In this paper, we introduce a new framework for unsupervised deep homography estimation. Our contributions are 3 folds. First, unlike previous methods that regr

ess 4 offsets for a homography, we propose a homography flow representation, whi ch can be estimated by a weighted sum of 8 pre-defined homography flow bases. Se cond, considering a homography contains 8 Degree-of-Freedoms (DOFs) that is much less than the rank of the network features, we propose a Low Rank Representatio n (LRR) block that reduces the feature rank, so that features corresponding to t he dominant motions are retained while others are rejected. Last, we propose a F eature Identity Loss (FIL) to enforce the learned image feature warp-equivariant , meaning that the result should be identical if the order of warp operation and feature extraction is swapped. With this constraint, the unsupervised optimizat ion is achieved more effectively and more stable features are learned. Extensive experiments are conducted to demonstrate the effectiveness of all the newly pro posed components, and results show that our approach outperforms the state-of-th e-art on the homography benchmark datasets both qualitatively and quantitatively . Code is available at https://github.com/megvii-research/BasesHomo

****

DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras

Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, Yebi n Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6239-6249

We propose DeepMultiCap, a novel method for multi-person performance capture usi ng sparse multi-view cameras. Our method can capture time varying surface detail s without the need of using pre-scanned template models. To tackle with the seri ous occlusion challenge for close interacting scenes, we combine a recently prop osed pixel-aligned implicit function with parametric model for robust reconstruc tion of the invisible surface areas. An effective attention-aware module is desi gned to obtain the fine-grained geometry details from multi-view images, where h igh-fidelity results can be generated. In addition to the spatial attention meth od, for video inputs, we further propose a novel temporal fusion method to allev iate the noise and temporal inconsistencies for moving character reconstruction. For quantitative evaluation, we contribute a high quality multi-person dataset, MultiHuman, which consists of 150 static scenes with different levels of occlus ions and ground truth 3D human models. Experimental results demonstrate the stat e-of-the-art performance of our method and the well generalization to real multi view video data, which outperforms the prior works by a large margin.

****

AI Choreographer: Music Conditioned 3D Dance Generation With AIST++

Ruilong Li, Shan Yang, David A. Ross, Angjoo Kanazawa; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 13401-13412

We present AIST++, a new multi-modal dataset of 3D dance motion and music, along with FACT, a Full-Attention Cross-modal Transformer network for generating 3D d ance motion conditioned on music. The proposed AIST++ dataset contains 1.1M fram es of 3D dance motion in 1408 sequences, covering 10 dance genres with multi-vie w videos with known camera poses---the largest dataset of this kind to our knowl edge. We show that naively applying sequence models such as transformers to this dataset for the task of music conditioned 3D motion generation does not produce satisfactory 3D motion that is well correlated with the input music. We overcom e these shortcomings by introducing key changes in its architecture design and s upervision: FACT model involves a deep cross-modal transformer block with full-a ttention that is trained to predict N future motions. We empirically show that t hese changes are key factors in generating long sequences of realistic dance mot ion that are well-attuned to the input music. We conduct extensive experiments o n AIST++ with user studies, where our method outperforms recent state-of-the-art methods both qualitatively and quantitatively. The code and the dataset can be found at: https://google.github.io/aichoreographer.

****

PU-EVA: An Edge-Vector Based Approximation Solution for Flexible-Scale Point Clo ud Upsampling

Luqing Luo, Lulu Tang, Wanyi Zhou, Shizheng Wang, Zhi-Xin Yang; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16208-

16217
High-quality point clouds have practical significance for point-based rendering, semantic understanding, and surface reconstruction. Upsampling sparse, noisy and non-uniform point clouds for a denser and more regular approximation of target objects is a desirable but challenging task. Most existing methods duplicate point features for upsampling, constraining the upsampling scales at a fixed rate. In this work, the arbitrary point clouds upsampling rates are achieved via edge-vector based affine combinations, and a novel design of Edge-Vector based Approximation for Flexible-scale Point clouds Upsampling (PU-EVA) is proposed. The edge-vector based approximation encodes neighboring connectivity via affine combinations based on edge vectors, and restricts the approximation error within a second-order term of Taylor's Expansion. Moreover, the EVA upsampling decouples the upsampling scales with network architecture, achieving the arbitrary upsampling rates in one-time training. Qualitative and quantitative evaluations demonstrate that the proposed PU-EVA outperforms the state-of-the-arts in terms of proximity-to-surface, distribution uniformity, and geometric details preservation.
****

Spatial Uncertainty-Aware Semi-Supervised Crowd Counting
Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Hung, Yalin Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15549-15559
Semi-supervised approaches for crowd counting attract attention, as the fully supervised paradigm is expensive and laborious due to its request for a large number of images of dense crowd scenarios and their annotations. This paper proposes a spatial uncertainty-aware semi-supervised approach via regularized surrogate task (binary segmentation) for crowd counting problems. Different from existing semi-supervised learning-based crowd counting methods, to exploit the unlabeled data, our proposed spatial uncertainty-aware teacher-student framework focuses on high confident regions' information while addressing the noisy supervision from the unlabeled data in an end-to-end manner. Specifically, we estimate the spatial uncertainty maps from the teacher model's surrogate task to guide the feature learning of the main task (density regression) and the surrogate task of the student model at the same time. Besides, we introduce a simple yet effective differential transformation layer to enforce the inherent spatial consistency regularization between the main task and the surrogate task in the student model, which helps the surrogate task to yield more reliable predictions and generates high-quality uncertainty maps. Thus, our model can also address the task-level perturbation problems that occur spatial inconsistency between the primary and surrogate tasks in the student model. Experimental results on four challenging crowd counting datasets demonstrate that our method achieves superior performance to the state-of-the-art semi-supervised methods. Code is available at : https://github.com/smallmax00/SUA_crowd_counting
****

SurfGen: Adversarial 3D Shape Synthesis With Explicit Surface Discriminators
Andrew Luo, Tianqin Li, Wen-Hao Zhang, Tai Sing Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16238-16248
Recent advances in deep generative models have led to immense progress in 3D shape synthesis. While existing models are able to synthesize shapes represented as voxels, point-clouds, or implicit functions, these methods only indirectly enforce the plausibility of the final 3D shape surface. Here we present a 3D shape synthesis framework (SurfGen) that directly applies adversarial training to the object surface. Our approach uses a differentiable spherical projection layer to capture and represent the explicit zero isosurface of an implicit 3D generator as functions defined on the unit sphere. By processing the spherical representation of 3D object surfaces with a spherical CNN in an adversarial setting, our generator can better learn the statistics of natural shape surfaces. We evaluate our model on large-scale shape datasets, and demonstrate that the end-to-end trained model is capable of generating high fidelity 3D shapes with diverse topology
****

TransReID: Transformer-Based Object Re-Identification

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, Wei Jiang; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15013-1 5022

Extracting robust feature representation is one of the key challenges in object re-identification (ReID). Although convolution neural network (CNN)-based method s have achieved great success, they only process one local neighborhood at a tim e and suffer from information loss on details caused by convolution and downsamp ling operators pooling and strided convolution).To overcome these limitations, w e propose a pure transformer-based object ReID framework named TransReID. Specif ically, we first encode an image as a sequence of patches and build a transforme r-based strong baseline with a few critical improvements, which achieves competi tive results on several ReID benchmarks with CNN-based methods. To further enhan ce the robust feature learning in the context of transformers, two novel modules are carefully designed. (i) The jigsaw patch module (JPM) is proposed to rearra nge the patch embeddings via shift and patch shuffle operations which generates robust features with improved discrimination ability and more diversified covera ge. (ii) The side information embeddings (SIE) is introduced to mitigate feature bias towards camera/view variations by plugging in learnable embeddings to inco rporate these non-visual clues. To the best of our knowledge, this is the first work to adopt a pure transformer for ReID research. Experimental results of Tran sReID are superior promising, which achieve state-of-the-art performance on both person and vehicle ReID benchmarks. Code is available at https://github.com/hes huting555/TransReID

****

Batch Normalization Increases Adversarial Vulnerability and Decreases Adversaria l Transferability: A Non-Robust Feature Perspective

Philipp Benz, Chaoning Zhang, In So Kweon; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 7818-7827

Batch normalization (BN) has been widely used in modern deep neural networks (DN Ns) due to improved convergence. BN is observed to increase the model accuracy w hile at the cost of adversarial robustness. There is an increasing interest in t he ML community to understand the impact of BN on DNNs, especially related to th e model robustness. This work attempts to understand the impact of BN on DNNs fr om a non-robust feature perspective. Straightforwardly, the improved accuracy ca n be attributed to the better utilization of useful features. It remains unclear whether BN mainly favors learning robust features (RFs) or non-robust features (NRFs). Our work presents empirical evidence that supports that BN shifts a mode l towards being more dependent on NRFs. To facilitate the analysis of such a fea ture robustness shift, we propose a framework for disentangling robust usefulnes s into robustness and usefulness. Extensive analysis under the proposed framewor k yields valuable insight on the DNN behavior regarding robustness, e.g. DNNs fi rst mainly learn RFs and then NRFs. The insight that RFs transfer better than NR Fs, further inspires simple techniques to strengthen transfer-based black-box at tacks.

****

Foreground Activation Maps for Weakly Supervised Object Localization

Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, Feng Wu; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3385-3395

Weakly supervised object localization (WSOL) aims to localize objects with only image-level labels, which has better scalability and practicability than fully s upervised methods in the actual deployment. However, with only image-level label s, learning object classification models tends to activate object parts and igno re the whole object, while expanding object parts into the whole object may dete riorate classification performance. To alleviate this problem, we propose foregr ound activation maps (FAM), whose aim is to optimize object localization and cla ssification jointly via an object-aware attention module and a part-aware attent ion module in a unified model, where the two tasks can complement and enhance ea ch other. To the best of our knowledge, this is the first work that can achieve remarkable performance for both tasks by optimizing them jointly via FAM for WSO L. Besides, the designed two modules can effectively highlight foreground object

s for localization and discover discriminative parts for classification. Extensive experiments with four backbones on two standard benchmarks demonstrate that our FAM performs favorably against state-of-the-art WSOL methods.
****

Self-Mutual Distillation Learning for Continuous Sign Language Recognition
Aiming Hao, Yuecong Min, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11303-11312
In recent years, deep learning moves video-based Continuous Sign Language Recognition (CSLR) significantly forward. Currently, a typical network combination for CSLR includes a visual module, which focuses on spatial and short-temporal information, followed by a contextual module, which focuses on long-temporal information, and the Connectionist Temporal Classification (CTC) loss is adopted to train the network. However, due to the limitation of chain rules in back-propagation, the visual module is hard to adjust for seeking optimized visual features. As a result, it enforces that the contextual module focuses on contextual information optimization only rather than balancing efficient visual and contextual information. In this paper, we propose a Self-Mutual Knowledge Distillation (SMKD) method, which enforces the visual and contextual modules to focus on short-term and long-term information and enhances the discriminative power of both modules simultaneously. Specifically, the visual and contextual modules share the weights of their corresponding classifiers, and train with CTC loss simultaneously. Moreover, the spike phenomenon widely exists with CTC loss. Although it can help us choose a few of the key frames of a gloss, it does drop other frames in a gloss and makes the visual feature saturation in the early stage. A gloss segmentation is developed to relieve the spike phenomenon and decrease saturation in the visual module. We conduct experiments on two CSLR benchmarks: PHOENIX14 and PHOENIX14-T. Experimental results demonstrate the effectiveness of the SMKD.
****

SOMA: Solving Optical Marker-Based MoCap Automatically
Nima Ghorbani, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11117-11126
Marker-based optical motion capture (mocap) is the "gold standard" method for acquiring accurate 3D human motion in computer vision, medicine, and graphics. The raw output of these systems are noisy and incomplete 3D points or short tracklets of points. To be useful, one must associate these points with corresponding markers on the captured subject; i.e. "labelling". Given these labels, one can then "solve" for the 3D skeleton or body surface mesh. Commercial auto-labeling tools require a specific calibration procedure at capture time, which is not possible for archival data. Here we train a novel neural network called SOMA, which takes raw mocap point clouds with varying numbers of points, labels them at scale without any calibration data, independent of the capture technology, and requiring only minimal human intervention. Our key insight is that, while labeling point clouds is highly ambiguous, the 3D body provides strong constraints on the solution that can be exploited by a learning-based method. To enable learning, we generate massive training sets of simulated noisy and ground truth mocap markers animated by 3D bodies from AMASS. SOMA exploits an architecture with stacked self-attention elements to learn the spatial structure of the 3D body and an optimal transport layer to constrain the assignment (labeling) problem while rejecting outliers. We extensively evaluate SOMA both quantitatively and qualitatively. SOMA is more accurate and robust than existing state of the art research methods and can be applied where commercial systems cannot. We automatically label over 8 hours of archival mocap data across 4 different datasets captured using various technologies and output SMPL-X body models. The model and data is released for research purposes at https://soma.is.tue.mpg.de/.
****

GLiT: Neural Architecture Search for Global and Local Image Transformer
Boyu Chen, Peixia Li, Chuming Li, Baopu Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12-21
We introduce the first Neural Architecture Search (NAS) method to find a better

transformer architecture for image recognition. Recently, transformers without CNN-based backbones are found to achieve impressive performance for image recognition. However, the transformer is designed for NLP tasks and thus could be sub-optimal when directly used for image recognition. In order to improve the visual representation ability for transformers, we propose a new search space and searching algorithm. Specifically, we introduce a locality module that models the local correlations in images explicitly with fewer computational cost. With the locality module, our search space is defined to let the search algorithm freely trade off between global and local information as well as optimizing the low-level design choice in each module. To tackle the problem caused by huge search space, a hierarchical neural architecture search method is proposed to search the optimal vision transformer from two levels separately with the evolutionary algorithm. Extensive experiments on the ImageNet dataset demonstrate that our method can find more discriminative and efficient transformer variants than the ResNet family (e.g., ResNet101) and the baseline ViT for image classification.
****

In-the-Wild Single Camera 3D Reconstruction Through Moving Water Surfaces
Jinhui Xiong, Wolfgang Heidrich; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12558-12567
We present a method for reconstructing the 3D shape of underwater environments from a single, stationary camera placed above the water. We propose a novel differentiable framework, which, to our knowledge, is the first single-camera solution that is capable of simultaneously retrieving the structure of dynamic water surfaces and static underwater scene geometry in the wild. This framework integrates ray casting of Snell's law at the refractive interface, multi-view triangulation and specially designed loss functions. Our method is calibration-free, and thus it is easy to collect data outdoors in uncontrolled environments. Experimental results show that our method is able to realize robust and quality reconstructions on a variety of scenes, both in a laboratory environment and in the wild, and even in a salt water environment. We believe the method is promising for applications in surveying and environmental monitoring.
****

DeepCAD: A Deep Generative Network for Computer-Aided Design Models
Rundi Wu, Chang Xiao, Changxi Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6772-6782
Deep generative models of 3D shapes have received a great deal of research interest. Yet, almost all of them generate discrete shape representations, such as voxels, point clouds, and polygon meshes. We present the first 3D generative model for a drastically different shape representation --- describing a shape as a sequence of computer-aided design (CAD) operations. Unlike meshes and point clouds, CAD models encode the user creation process of 3D shapes, widely used in numerous industrial and engineering design tasks. However, the sequential and irregular structure of CAD operations poses significant challenges for existing 3D generative models. Drawing an analogy between CAD operations and natural language, we propose a CAD generative network based on the Transformer. We demonstrate the performance of our model for both shape autoencoding and random shape generation. To train our network, we create a new CAD dataset consisting of 178,238 models and their CAD construction sequences. We have made this dataset publicly available to promote future research on this topic.
****

Understanding Robustness of Transformers for Image Classification
Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, Andreas Veit; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10231-10241
Deep Convolutional Neural Networks (CNNs) have long been the architecture of choice for computer vision tasks. Recently, Transformer-based architectures like Vision Transformer (ViT) have matched or even surpassed ResNets for image classification. However, details of the Transformer architecture such as the use of non-overlapping patches lead one to wonder whether these networks are as robust. In this paper, we perform an extensive study of a variety of different measures of

robustness of ViT models and compare the findings to ResNet baselines. We investigate robustness to input perturbations as well as robustness to model perturbations. We find that when pre-trained with a sufficient amount of data, ViT models are at least as robust as the ResNet counterparts on a broad range of perturbations. We also find that Transformers are robust to the removal of almost any single layer, and that while activations from later layers are highly correlated with each other, they nevertheless play an important role in classification.
****

Learning Canonical View Representation for 3D Shape Recognition With Arbitrary Views
Xin Wei, Yifei Gong, Fudong Wang, Xing Sun, Jian Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 407-416
In this paper, we focus on recognizing 3D shapes from arbitrary views, i.e., arbitrary numbers and positions of viewpoints. It is a challenging and realistic setting for view-based 3D shape recognition. We propose a canonical view representation to tackle this challenge. We first transform the original features of arbitrary views to a fixed number of view features, dubbed canonical view representation, by aligning the arbitrary view features to a set of learnable reference view features using optimal transport. In this way, each 3D shape with arbitrary views is represented by a fixed number of canonical view features, which are further aggregated to generate a rich and robust 3D shape representation for shape recognition. We also propose a canonical view feature separation constraint to enforce that the view features in canonical view representation can be embedded into scattered points in a Euclidean space. Experiments on the ModelNet40, ScanObjectNN, and RGBD datasets show that our method achieves competitive results under the fixed viewpoint settings, and significantly outperforms the applicable methods under the arbitrary view setting.
****

Fourier Space Losses for Efficient Perceptual Image Super-Resolution
Dario Fuoli, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2360-2369
Many super-resolution (SR) models are optimized for high performance only and therefore lack efficiency due to large model complexity. As large models are often not practical in real-world applications, we investigate and propose novel loss functions, to enable SR with high perceptual quality from much more efficient models. The representative power for a given low-complexity generator network can only be fully leveraged by strong guidance towards the optimal set of parameters. We show that it is possible to improve the performance of a recently introduced efficient generator architecture solely with the application of our proposed loss functions. In particular, we use a Fourier space supervision loss for improved restoration of missing high-frequency (HF) content from the ground truth image and design a discriminator architecture working directly in the Fourier domain to better match the target HF distribution. We show that our losses' direct emphasis on the frequencies in Fourier-space significantly boosts the perceptual image quality, while at the same time retaining high restoration quality in comparison to previously proposed loss functions for this task. The performance is further improved by utilizing a combination of spatial and frequency domain losses, as both representations provide complementary information during training. On top of that, the trained generator achieves comparable results with and is 2.4x and 48x faster than state-of-the-art perceptual SR methods RankSRGAN and SRFlow respectively.
****

A Backdoor Attack Against 3D Point Cloud Classifiers
Zhen Xiang, David J. Miller, Siheng Chen, Xi Li, George Kesidis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7597-7607
Vulnerability of 3D point cloud (PC) classifiers has become a grave concern due to the popularity of 3D sensors in safety-critical applications. Existing adversarial attacks against 3D PC classifiers are all test-time evasion (TTE) attacks that aim to induce test-time misclassifications using knowledge of the classifie

r. But since the victim classifier is usually not accessible to the attacker, th e threat is largely diminished in practice, as PC TTEs typically have poor trans ferability. Here, we propose the first backdoor attack (BA) against PC classifie rs. Originally proposed for images, BAs poison the victim classifier's training set so that the classifier learns to decide to the attacker's target class whene ver the attacker's backdoor pattern is present in a given input sample. Signific antly, BAs do not require knowledge of the victim classifier. Different from ima ge BAs, we propose to insert a cluster of points into a PC as a robust backdoor pattern customized for 3D PCs. Such clusters are also consistent with a physical attack (i.e., with a captured object in a scene). We optimize the cluster's loc ation using an independently trained surrogate classifier and choose the cluster 's local geometry to evade possible PC preprocessing and PC anomaly detectors (A Ds). Experimentally, our BA achieves a uniformly high success rate (>=87%) and s hows evasiveness against state-of-the-art PC ADs. Code is available at https://g ithub.com/zhenxianglance/PCBA.

****

Guided Point Contrastive Learning for Semi-Supervised Point Cloud Semantic Segme ntation

Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6423-6432

Rapid progress in 3D semantic segmentation is inseparable from the advances of d eep network models, which highly rely on large-scale annotated data for training . To address the high cost and challenges of 3D point-level labeling, we present a method for semi-supervised point cloud semantic segmentation to adopt unlabel ed point clouds in training to boost the model performance. Inspired by the rece nt contrastive loss in self-supervised tasks, we propose the guided point contra stive loss to enhance the feature representation and model generalization abilit y in semi-supervised setting. Semantic predictions on unlabeled point clouds ser ve as pseudo-label guidance in our loss to avoid negative pairs in the same cate gory. Also, we design the confidence guidance to ensure high-quality feature lea rning. Besides, a category-balanced sampling strategy is proposed to collect pos itive and negative samples to mitigate the class imbalance problem. Extensive ex periments on three datasets (ScanNet V2, S3DIS, and SemanticKITTI) show the effe ctiveness of our semi-supervised method to improve the prediction quality with u nlabeled data.

****

Location-Aware Single Image Reflection Removal

Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 50 17-5026

This paper proposes a novel location-aware deep-learning-based single image refl ection removal method. Our network has a reflection detection module to regress a probabilistic reflection confidence map, taking multi-scale Laplacian features as inputs. This probabilistic map tells if a region is reflection-dominated or transmission-dominated, and it is used as a cue for the network to control the f eature flow when predicting the reflection and transmission layers. We design ou r network as a recurrent network to progressively refine reflection removal resu lts at each iteration. The novelty is that we leverage Laplacian kernel paramete rs to emphasize the boundaries of strong reflections. It is beneficial to strong reflection detection and substantially improves the quality of reflection remov al results. Extensive experiments verify the superior performance of the propose d method over state-of-the-art approaches. Our code and the pre-trained model ca n be found at https://github.com/zdlarr/Location-aware-SIRR.

****

Better Aggregation in Test-Time Augmentation

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, John Guttag; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1214-1 223

Test-time augmentation---the aggregation of predictions across transformed versi

ons of a test input---is a common practice in image classification. Traditionally, predictions are combined using a simple average. In this paper, we present 1) experimental analyses that shed light on cases in which the simple average is suboptimal and 2) a method to address these shortcomings. A key finding is that even when test-time augmentation produces a net improvement in accuracy, it can change many correct predictions into incorrect predictions. We delve into when and why test-time augmentation changes a prediction from being correct to incorrect and vice versa. Building on these insights, we present a learning-based method for aggregating test-time augmentations. Experiments across a diverse set of models, datasets, and augmentations show that our method delivers consistent improvements over existing approaches.

****

Self-Born Wiring for Neural Trees

Ying Chen, Feng Mao, Jie Song, Xinchao Wang, Huiqiong Wang, Mingli Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5047-5056

Neural trees aim at integrating deep neural networks and decision trees so as to bring the best of the two worlds, including representation learning from the former and faster inference from the latter. In this paper, we introduce a novel approach, termed as Self-born Wiring (SeBoW), to learn neural trees from a mother deep neural network. In contrast to prior neural-tree approaches that either adopt a pre-defined structure or grow hierarchical layers in a progressive manner, task-adaptive neural trees in SeBoW evolve from a deep neural network through a construction-by-destruction process, enabling a global-level parameter optimization that further yields favorable results. Specifically, given a designated network configuration like VGG, SeBoW disconnects all the layers and derives isolated filter groups, based on which a global-level wiring process is conducted to attach a subset of filter groups, eventually bearing a lightweight neural tree. Extensive experiments demonstrate that, with a lower computational cost, SeBoW outperforms all prior neural trees by a significant margin and even achieves results on par with predominant non-tree networks like ResNets. Moreover, SeBoW proves its scalability to large-scale datasets like ImageNet, which has been barely explored by prior tree networks.

****

DenseTNT: End-to-End Trajectory Prediction From Dense Goal Sets

Junru Gu, Chen Sun, Hang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15303-15312

Due to the stochasticity of human behaviors, predicting the future trajectories of road agents is challenging for autonomous driving. Recently, goal-based multi-trajectory prediction methods are proved to be effective, where they first score over-sampled goal candidates and then select a final set from them. However, these methods usually involve goal predictions based on sparse pre-defined anchors and heuristic goal selection algorithms. In this work, we propose an anchor-free and end-to-end trajectory prediction model, named DenseTNT, that directly outputs a set of trajectories from dense goal candidates. In addition, we introduce an offline optimization-based technique to provide multi-future pseudo-labels for our final online model. Experiments show that DenseTNT achieves state-of-the-art performance, ranking 1st on the Argoverse motion forecasting benchmark and being the 1st place winner of the 2021 Waymo Open Dataset Motion Prediction Challenge.

****

Segmentation-Grounded Scene Graph Generation

Siddhesh Khandelwal, Mohammed Suhail, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15879-15889

Scene graph generation has emerged as an important problem in computer vision. While scene graphs provide a grounded representation of objects, their locations and relations in an image, they do so only at the granularity of proposal bounding boxes. In this work, we propose the first, to our knowledge, framework for pixel-level segmentation-grounded scene graph generation. Our framework is agnostic to the underlying scene graph generation method and address the lack of segmen

tation annotations in target scene graph datasets (e.g., Visual Genome) through transfer and multi-task learning from, and with, an auxiliary dataset (e.g., MS COCO). Specifically, each target object being detected is endowed with a segmentation mask, which is expressed as a lingual-similarity weighted linear combination over categories that have annotations present in an auxiliary dataset. These inferred masks, along with a Gaussian masking mechanism which grounds the relations at a pixel-level within the image, allow for improved relation prediction. The entire framework is end-to-end trainable and is learned in a multi-task manner.

****

Detector-Free Weakly Supervised Grounding by Separation

Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, Chun-Fu (Richard) Chen, Alex Bronstein, Kate Saenko, Shimon Ullman, Raja Giryes, Rogerio Feris, Leonid Karlinsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1801-1812

Nowadays, there is an abundance of data involving images and surrounding free-form text weakly corresponding to those images. Weakly Supervised phrase-Grounding (WSG) deals with the task of using this data to learn to localize (or to ground) arbitrary text phrases in images without any additional annotations. However, most recent SotA methods for WSG assume an existence of a pre-trained object detector, relying on it to produce the ROIs for localization. In this work, we focus on the task of Detector-Free WSG (DF-WSG) to solve WSG without relying on a pre-trained detector. We directly learn everything from the images and associated free-form text pairs, thus potentially gaining advantage on the categories unsupported by the detector. The key idea behind our proposed Grounding by Separation (GbS) method is synthesizing `text to image-regions' associations by random alpha-blending of arbitrary image pairs and using the corresponding texts of the pair as conditions to recover the alpha map from the blended image via a segmentation network. At test time, this allows using the query phrase as a condition for a non-blended query image, thus interpreting the test image as a composition of a region corresponding to the phrase and the complement region. Using this approach we demonstrate a significant accuracy improvement, up to 8.5% over previous DF-WSG SotA, for a range of benchmarks including Flickr30K, Visual Genome, and ReferIt, as well as a significant complementary improvement (above 7%) over the detector-based approaches for WSG.

****

Geography-Aware Self-Supervised Learning

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, Stefano Ermon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10181-10190

Contrastive learning methods have significantly narrowed the gap between supervised and unsupervised learning on computer vision tasks. In this paper, we explore their application to geo-located datasets, e.g. remote sensing, where unlabeled data is often abundant but labeled data is scarce. We first show that due to their different characteristics, a non-trivial gap persists between contrastive and supervised learning on standard benchmarks. To close the gap, we propose novel training methods that exploit the spatio-temporal structure of remote sensing data. We leverage spatially aligned images over time to construct temporal positive pairs in contrastive learning and geo-location to design pre-text tasks. Our experiments show that our proposed method closes the gap between contrastive and supervised learning on image classification, object detection and semantic segmentation for remote sensing. Moreover, we demonstrate that the proposed method can also be applied to geo-tagged ImageNet images, improving downstream performance on various tasks.

****

CrossCLR: Cross-Modal Contrastive Learning for Multi-Modal Video Representations

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, Thomas Brox; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1450-1459

Contrastive learning allows us to flexibly define powerful losses by contrasting

positive pairs from sets of negative samples. Recently, the principle has also been used to learn cross-modal embeddings for video and text, yet without exploiting its full potential. In particular, previous losses do not take the intra-modality similarities into account, which leads to inefficient embeddings, as the same content is mapped to multiple points in the embedding space. With CrossCLR, we present a contrastive loss that fixes this issue. Moreover, we define sets of highly related samples in terms of their input embeddings and exclude them from the negative samples to avoid issues with false negatives. We show that these principles consistently improve the quality of the learned embeddings. The joint embeddings learned with CrossCLR extend the state of the art in video-text retrieval on Youcook2 and LSMDC datasets and in video captioning on the Youcook2 dataset by a large margin. We also demonstrate the generality of the concept by learning improved joint embeddings for other pairs of modalities.
****

Shape-Aware Multi-Person Pose Estimation From Multi-View Images
Zijian Dong, Jie Song, Xu Chen, Chen Guo, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11158-11168
In this paper we contribute a simple yet effective approach for estimating 3D poses of multiple people from multi-view images. Our proposed coarse-to-fine pipeline first aggregates noisy 2D observations from multiple camera views into 3D space and then associates them into individual instances based on a confidence-aware majority voting technique. The final pose estimates are attained from a novel optimization scheme which links high-confidence multi-view 2D observations and 3D joint candidates. Moreover, a statistical parametric body model such as SMPL is leveraged as a regularizing prior for these 3D joint candidates. Specifically, both 3D poses and SMPL parameters are optimized jointly in an alternating fashion. Here the parametric models help in correcting implausible 3D pose estimates and filling in missing joint detections while updated 3D poses in turn guide obtaining better SMPL estimations. By linking 2D and 3D observations, our method is both accurate and generalizes to different data sources because it better decouples the final 3D pose from the inter-person constellation and is more robust to noisy 2D detections. We systematically evaluate our method on public datasets and achieve state-of-the-art performance. The code and video will be available on the project page: https://ait.ethz.ch/projects/2021/multi-human-pose/.
****

Single Image Defocus Deblurring Using Kernel-Sharing Parallel Atrous Convolutions
Hyeongseok Son, Junyong Lee, Sunghyun Cho, Seungyong Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2642-2650
This paper proposes a novel deep learning approach for single image defocus deblurring based on inverse kernels. In a defocused image, the blur shapes are similar among pixels although the blur sizes can spatially vary. To utilize the property with inverse kernels, we exploit the observation that when only the size of a defocus blur changes while keeping the shape, the shape of the corresponding inverse kernel remains the same and only the scale changes. Based on the observation, we propose a kernel-sharing parallel atrous convolutional (KPAC) block specifically designed by incorporating the property of inverse kernels for single image defocus deblurring. To effectively simulate the invariant shapes of inverse kernels with different scales, KPAC shares the same convolutional weights among multiple atrous convolution layers. To efficiently simulate the varying scales of inverse kernels, KPAC consists of only a few atrous convolution layers with different dilations and learns per-pixel scale attentions to aggregate the outputs of the layers. KPAC also utilizes the shape attention to combine the outputs of multiple convolution filters in each atrous convolution layer, to deal with defocus blur with a slightly varying shape. We demonstrate that our approach achieves state-of-the-art performance with a much smaller number of parameters than previous methods.
****

Time-Multiplexed Coded Aperture Imaging: Learned Coded Aperture and Pixel Exposures for Compressive Imaging Systems

Edwin Vargas, Julien N. P. Martel, Gordon Wetzstein, Henry Arguello; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2692-2702

Compressive imaging using coded apertures (CA) is a powerful technique that can be used to recover depth, light fields, hyperspectral images and other quantities from a single snapshot. The performance of compressive imaging systems based on CAs mostly depends on two factors: the properties of the mask's attenuation pattern, that we refer to as "codification", and the computational techniques used to recover the quantity of interest from the coded snapshot. In this work, we introduce the idea of using time-varying CAs synchronized with spatially varying pixel shutters. We divide the exposure of a sensor into sub-exposures at the beginning of which the CA mask changes and at which the sensor's pixels are simultaneously and individually switched "on" or "off". This is a practically appealing codification as it does not introduce additional optical components other than the already present CA but uses a change in the pixel shutter that can be easily realized electronically. We show that our proposed time-multiplexed coded aperture (TMCA) can be optimized end to end and induces better coded snapshots enabling superior reconstructions in two different applications: compressive light field imaging and hyperspectral imaging. We demonstrate both in simulation and with real captures (taken with prototypes we built) that this codification outperforms the state-of-the-art compressive imaging systems by a large margin in those applications.

****

Motion-Aware Dynamic Architecture for Efficient Frame Interpolation

Myungsub Choi, Suyoung Lee, Heewon Kim, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13839-13848

Video frame interpolation aims to synthesize accurate intermediate frames given a low-frame-rate video. While the quality of the generated frames is increasingly getting better, state-of-the-art models have become more and more computationally expensive. However, local regions with small or no motion can be easily interpolated with simple models and do not require such heavy compute, whereas some regions may not be correct even after inference through a large model. Thus, we propose an effective framework that assigns varying amounts of computation for different regions. Our dynamic architecture first calculates the approximate motion magnitude to use as a proxy for the difficulty levels for each region, and decides the depth of the model and the scale of the input. Experimental results show that static regions pass through a smaller number of layers, while the regions with larger motion are downscaled for better motion reasoning. In doing so, we demonstrate that the proposed framework can significantly reduce the computation cost (FLOPs) while maintaining the performance, often up to 50% when interpolating a 2K resolution video.

****

Contrasting Contrastive Self-Supervised Representation Learning Pipelines

Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, Roozbeh Mottaghi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9949-9959

In the past few years, we have witnessed remarkable breakthroughs in self-supervised representation learning. Despite the success and adoption of representations learned through this paradigm, much is yet to be understood about how different training methods and datasets influence performance on downstream tasks. In this paper, we analyze contrastive approaches as one of the most successful and popular variants of self-supervised representation learning. We perform this analysis from the perspective of the training algorithms, pre-training datasets and end tasks. We examine over 700 training experiments including 30 encoders, 4 pre-training datasets and 20 diverse downstream tasks. Our experiments address various questions regarding the performance of self-supervised models compared to their supervised counterparts, current benchmarks used for evaluation, and the effect of the pre-training data on end task performance. We hope the insights and empirical evidence provided by this work will help future research in learning better visual representations.

****
Normalized Human Pose Features for Human Action Video Alignment
Jingyuan Liu, Mingyi Shi, Qifeng Chen, Hongbo Fu, Chiew-Lan Tai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11521-11531
We present a novel approach for extracting human pose features from human action videos. The goal is to let the pose features capture only the poses of the action while being invariant to other factors, including video backgrounds, the video subject's anthropometric characteristics and viewpoints. Such human pose features facilitate the comparison of pose similarity and can be used for down-stream tasks, such as human action video alignment and pose retrieval. The key to our approach is to first normalize the poses in the video frames by retargeting the poses onto a pre-defined 3D skeleton to not only disentangle subject physical features, such as bone lengths and ratios, but also to unify global orientations of the poses. Then the normalized poses are mapped to a pose embedding space of high-level features, learned via unsupervised metric learning. We evaluate the effectiveness of our normalized features both qualitatively by visualizations, and quantitatively by a video alignment task on the Human3.6M dataset and an action recognition task on the Penn Action dataset.
****
Learning Hierarchical Graph Neural Networks for Image Clustering
Yifan Xing, Tong He, Tianjun Xiao, Yongxin Wang, Yuanjun Xiong, Wei Xia, David Wipf, Zheng Zhang, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3467-3477
We propose a hierarchical graph neural network (GNN) model that learns how to cluster a set of images into an unknown number of identities using a training set of images annotated with labels belonging to a disjoint set of identities. Our hierarchical GNN uses a novel approach to merge connected components predicted at each level of the hierarchy to form a new graph at the next level. Unlike fully unsupervised hierarchical clustering, the choice of grouping and complexity criteria stems naturally from supervision in the training set. The resulting method, Hi-LANDER, achieves an average of 49% improvement in F-score and 7% increase in Normalized Mutual Information (NMI) relative to current GNN-based clustering algorithms. Additionally, state-of-the-art GNN-based methods rely on separate models to predict linkage probabilities and node densities as intermediate steps of the clustering process. In contrast, our unified framework achieves a three-fold decrease in computational cost. Our training and inference code are released.
****
Indoor Scene Generation From a Collection of Semantic-Segmented Depth Images
Ming-Jia Yang, Yu-Xiao Guo, Bin Zhou, Xin Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15203-15212
We present a method for creating 3D indoor scenes with a generative model learned from a collection of semantic-segmented depth images captured from different unknown scenes. Given a room with a specified size, our method automatically generates 3D objects in a room from a randomly sampled latent code. Different from existing methods that represent an indoor scene with the type, location, and other properties of objects in the room and learn the scene layout from a collection of complete 3D indoor scenes, our method models each indoor scene as a 3D semantic scene volume and learns a volumetric generative adversarial network (GAN) from a collection of 2.5D partial observations of 3D scenes. To this end, we apply a differentiable projection layer to project the generated 3D semantic scene volumes into semantic-segmented depth images and design a new multiple-view discriminator for learning the complete 3D scene volume from 2.5D semantic-segmented depth images. Compared to existing methods, our method not only efficiently reduces the workload of modeling and acquiring 3D scenes for training, but also produces better object shapes and their detailed layouts in the scene. We evaluate our method with different indoor scene datasets and demonstrate the advantages of our method. We also extend our method for generating 3D indoor scenes from semantic-segmented depth images inferred from RGB images of real scenes.
****

Keypoint Communities
Duncan Zauss, Sven Kreiss, Alexandre Alahi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11057-11066
We present a fast bottom-up method that jointly detects over 100 keypoints on humans or objects, also referred to as human/object pose estimation. We model all keypoints belonging to a human or an object --the pose-- as a graph and leverage insights from community detection to quantify the independence of keypoints. We use a graph centrality measure to assign training weights to different parts of a pose. Our proposed measure quantifies how tightly a keypoint is connected to its neighborhood. Our experiments show that our method outperforms all previous methods for human pose estimation with fine-grained keypoint annotations on the face, the hands and the feet with a total of 133 keypoints. We also show that our method generalizes to car poses.
****

Can Scale-Consistent Monocular Depth Be Learned in a Self-Supervised Scale-Invariant Manner?
Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12727-12736
Geometric constraints are shown to enforce scale consistency and remedy the scale ambiguity issue in self-supervised monocular depth estimation. Meanwhile, scale-invariant losses focus on learning relative depth, leading to accurate relative depth prediction. To combine the best of both worlds, we learn scale-consistent self-supervised depth in a scale-invariant manner. Towards this goal, we present a scale-aware geometric (SAG) loss, which enforces scale consistency through point cloud alignment. Compared to prior arts, SAG loss takes relative scale into consideration during relative motion estimation, enabling more precise alignment and explicit supervision for scale inference. In addition, a novel two-stream architecture for depth estimation is designed, which disentangles scale from depth estimation and allows depth to be learned in a scale-invariant manner. The integration of SAG loss and two-stream network enables more consistent scale inference and more accurate relative depth estimation. Our method achieves state-of-the-art performance under both scale-invariant and scale-dependent evaluation settings.
****

Multi-Task Self-Training for Learning General Representations
Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, Tsung-Yi Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8856-8865
Despite the fast progress in training specialized models for various tasks, learning a single general model that works well for many tasks is still challenging for computer vision. Here we introduce multi-task self-training (MuST), which harnesses the knowledge in independent specialized teacher models (e.g., ImageNet model on classification) to train a single general student model. Our approach has three steps. First, we train specialized teachers independently on labeled datasets. We then use the specialized teachers to label an unlabeled dataset to create a multi-task pseudo labeled dataset. Finally, the dataset, which now contains pseudo labels from teacher models trained on different datasets/tasks, is then used to train a student model with multi-task learning. We evaluate the feature representations of the student model on 6 vision tasks including image recognition (classification, detection, segmentation) and 3D geometry estimation (depth and surface normal estimation). MuST is scalable with unlabeled or partially labeled datasets and outperforms both specialized supervised models and self-supervised models when training on large scale datasets. Lastly, we show MuST can improve upon already strong checkpoints trained with billions of examples. The results suggest self-training is a promising direction to aggregate labeled and unlabeled training data for learning general feature representations.
****

Adaptive Unfolding Total Variation Network for Low-Light Image Enhancement
Chuanjun Zheng, Daming Shi, Wentian Shi; Proceedings of the IEEE/CVF Internation

al Conference on Computer Vision (ICCV), 2021, pp. 4439-4448
Real-world low-light images suffer from two main degradations, namely, inevitable noise and poor visibility. Since the noise exhibits different levels, its estimation has been implemented in recent works when enhancing low-light images from raw Bayer space. When it comes to sRGB color space, the noise estimation becomes more complicated due to the effect of the image processing pipeline. Nevertheless, most existing enhancing algorithms in sRGB space only focus on the low visibility problem or suppress the noise under a hypothetical noise level, leading them impractical due to the lack of robustness. To address this issue, we propose an adaptive unfolding total variation network (UTVNet), which approximates the noise level from the real sRGB low-light image by learning the balancing parameter in the model-based denoising method with total variation regularization. Meanwhile, we learn the noise level map by unrolling the corresponding minimization process for providing the inferences of smoothness and fidelity constraints. Guided by the noise level map, our UTVNet can recover finer details and is more capable to suppress noise in real captured low-light scenes. Extensive experiments on real-world low-light images clearly demonstrate the superior performance of UTVNet over state-of-the-art methods.
****

Training Weakly Supervised Video Frame Interpolation With Events
Zhiyang Yu, Yu Zhang, Deyuan Liu, Dongqing Zou, Xijun Chen, Yebin Liu, Jimmy S. Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14589-14598
Event-based video frame interpolation is promising as event cameras capture dense motion signals that can greatly facilitate motion-aware synthesis. However, training existing frameworks for this task requires high frame-rate videos with synchronized events, posing challenges to collect real training data. In this work we show event-based frame interpolation can be trained without the need of high framerate videos. This is achieved via a novel weakly supervised framework that 1) corrects image appearance by extracting complementary information from events and 2) supplants motion dynamics modeling with attention mechanisms. For the latter we propose subpixel attention learning, which supports searching high-resolution correspondence efficiently on low-resolution feature grid. Though trained on low frame-rate videos, our framework outperforms existing models trained with full high frame-rate videos (and events) on both GoPro dataset and a new real event-based dataset. Codes, models and dataset will be made available at: https://github.com/YU-Zhiyang/WEVI.
****

TransView: Inside, Outside, and Across the Cropping View Boundaries
Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, Weicai Zhong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4218-4227
We show that relation modeling between visual elements matters in cropping view recommendation. Cropping view recommendation addresses the problem of image recomposition conditioned on the composition quality and the ranking of views (cropped sub-regions). This task is challenging because the visual difference is subtle when a visual element is reserved or removed. Existing methods represent visual elements by extracting region-based convolutional features inside and outside the cropping view boundaries, without probing a fundamental question: why some visual elements are of interest or of discard? In this work, we observe that the relation between different visual elements significantly affects their relative positions to the desired cropping view, and such relation can be characterized by the attraction inside/outside the cropping view boundaries and the repulsion across the boundaries. By instantiating a transformer-based solution that represents visual elements as visual words and that models the dependencies between visual words, we report not only state of-the-art performance on public benchmarks, but also interesting visualizations that depict the attraction and repulsion between visual elements, which may shed light on what makes for effective cropping view recommendation.
****

Vis2Mesh: Efficient Mesh Reconstruction From Unstructured Point Clouds of Large

Scenes With Learned Virtual View Visibility
Shuang Song, Zhaopeng Cui, Rongjun Qin; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2021, pp. 6514-6524
We present a novel framework for mesh reconstruction from unstructured point clo uds by taking advantage of the learned visibility of the 3D points in the virtua l views and traditional graph-cut based mesh generation. Specifically, we first propose a three-step network that explicitly employs depth completion for visibi lity prediction. Then the visibility information of multiple views is aggregated to generate a 3D mesh model by solving an optimization problem considering visi bility in which a novel adaptive visibility weighting term in surface determinat ion is also introduced to suppress line of sight with a large incident angle. Co mpared to other learning-based approaches, our pipeline only exercises the learn ing on a 2D binary classification task, i.e., points visible or not in a view, w hich is much more generalizable and practically more efficient and capable to de al with a large number of points. Experiments demonstrate that our method with f avorable transferability and robustness, and achieve competing performances w.r. t. state-of-the-art learning-based approaches on small complex objects and outpe rforms on large indoor and outdoor scenes. Code is available at https://github.c om/GDAOSU/vis2mesh.
****

ID-Reveal: Identity-Aware DeepFake Video Detection
Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, Luisa Verdoli va; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 15108-15117
A major challenge in DeepFake forgery detection is that state-of-the-art algorit hms are mostly trained to detect a specific fake method. As a result, these appr oaches show poor generalization across different types of facial manipulations, e.g., from face swapping to facial reenactment. To this end, we introduce ID-Rev eal, a new approach that learns temporal facial features, specific of how a pers on moves while talking, by means of metric learning coupled with an adversarial training strategy. The advantage is that we do not need any training data of fak es, but only train on real videos. Moreover, we utilize high-level semantic feat ures, which enables robustess to widespread and disruptive forms of post-process ing. We perform a thorough experimental analysis on several publicly available b enchmarks. Compared to state of the art, our method improves generalization and is more robust to low-quality videos, that are usually spread over social networ ks. In particular, we obtain an average improvement of more than 15% in terms of accuracy for facial reenactment on high compressed videos.
****

GAN-Control: Explicitly Controllable GANs
Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, Gérard Medioni; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14083-1 4093
We present a framework for training GANs with explicit control over generated fa cial images. We are able to control the generated image by settings exact attrib utes such as age, pose, expression, etc. Most approaches for manipulating GAN-ge nerated images achieve partial control by leveraging the latent space disentangl ement properties, obtained implicitly after standard GAN training. Such methods are able to change the relative intensity of certain attributes, but not explici tly set their values. Recently proposed methods, designed for explicit control o ver human faces, harness morphable 3D face models (3DMM) to allow fine-grained c ontrol capabilities in GANs. Unlike these methods, our control is not constraine d to 3DMM parameters and is extendable beyond the domain of human faces. Using c ontrastive learning, we obtain GANs with an explicitly disentangled latent space . This disentanglement is utilized to train control-encoders mapping human-inter pretable inputs to suitable latent vectors, thus allowing explicit control. In t he domain of human faces we demonstrate control over identity, age, pose, expres sion, hair color and illumination. We also demonstrate control capabilities of o ur framework in the domains of painted portraits and dog image generation. We de monstrate that our approach achieves state-of-the-art performance both qualitati

vely and quantitatively.
****
A Closer Look at Rotation-Invariant Deep Point Cloud Analysis
Feiran Li, Kent Fujiwara, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16218-16227
We consider the deep point cloud analysis tasks where the inputs of the networks are randomly rotated. Recent progress in rotation-invariant point cloud analysis is mainly driven by converting point clouds into their respective canonical poses, and principal component analysis (PCA) is a practical tool to achieve this. Due to the imperfect alignment of PCA, most of the current works are devoted to developing powerful network structures and features to overcome this deficiency, without thoroughly analyzing the PCA-based canonical poses themselves. In this work, we present a detailed study w.r.t. the PCA-based canonical poses of point clouds. Our investigation reveals that the ambiguity problem associated with the PCA-based canonical poses is handled insufficiently in some recent works. To this end, we develop a simple pose selector module for disambiguation, which presents noticeable enhancement (i.e., 5:3% classification accuracy) over state-of-the-art approaches on the challenging real-world dataset.
****
Relating Adversarially Robust Generalization to Flat Minima
David Stutz, Matthias Hein, Bernt Schiele; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7807-7817
Adversarial training (AT) has become the de-facto standard to obtain models robust against adversarial examples. However, AT exhibits severe robust overfitting: cross-entropy loss on adversarial examples, so-called robust loss, decreases continuously on training examples, while eventually increasing on test examples. In practice, this leads to poor robust generalization, i.e., adversarial robustness does not generalize well to new examples. In this paper, we study the relationship between robust generalization and flatness of the robust loss landscape in weight space, i.e., whether robust loss changes significantly when perturbing weights. To this end, we propose average- and worst-case metrics to measure flatness in the robust loss landscape and show a correlation between good robust generalization and flatness. For example, throughout training, flatness reduces significantly during overfitting such that early stopping effectively finds flatter minima in the robust loss landscape. Similarly, AT variants achieving higher adversarial robustness also correspond to flatter minima. This holds for many popular choices, e.g., AT-AWP, TRADES, MART, AT with self-supervision or additional unlabeled examples, as well as simple regularization techniques, e.g., AutoAugment, weight decay or label noise. For fair comparison across these approaches, our flatness measures are specifically designed to be scale-invariant and we conduct extensive experiments to validate our findings.
****
Re-Energizing Domain Discriminator With Sample Relabeling for Adversarial Domain Adaptation
Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9174-9183
Many unsupervised domain adaptation (UDA) methods exploit domain adversarial training to align the features to reduce domain gap, where a feature extractor is trained to fool a domain discriminator in order to have aligned feature distributions. The discrimination capability of the domain classifier w.r.t. the increasingly aligned feature distributions deteriorates as training goes on, thus cannot effectively further drive the training of feature extractor. In this work, we propose an efficient optimization strategy named Re-enforceable Adversarial Domain Adaptation (RADA) which aims to re-energize the domain discriminator during the training by using dynamic domain labels. Particularly, we relabel the well aligned target domain samples as source domain samples on the fly. Such relabeling makes the less separable distributions more separable, and thus leads to a more powerful domain classifier w.r.t. the new data distributions, which in turn further drives feature alignment. Extensive experiments on multiple UDA benchmarks d

emonstrate the effectiveness and superiority of our RADA.
****
Learning To Adversarially Blur Visual Object Tracking
Qing Guo, Ziyi Cheng, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yang Liu, Jianjun Zh ao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 10839-10848
Motion blur caused by the moving of the object or camera during the exposure can be a key challenge for visual object tracking, affecting tracking accuracy sign ificantly. In this work, we explore the robustness of visual object trackers aga inst motion blur from a new angle, i.e., adversarial blur attack (ABA). Our main objective is to online transfer input frames to their natural motion-blurred co unterparts while misleading the state-of-the-art trackers during the tracking pr ocess. To this end, we first design the motion blur synthesizing method for visu al tracking based on the generation principle of motion blur, considering the mo tion information and the light accumulation process. With this synthetic method, we propose optimization-based ABA (OP-ABA) by iteratively optimizing an adversa rial objective function against the tracking w.r.t. the motion and light accumul ation parameters. The OP-ABA is able to produce natural adversarial examples but the iteration can cause heavy time cost, making it unsuitable for attacking rea l-time trackers. To alleviate this issue, we further propose one-step ABA (OS-AB A) where we design and train a joint adversarial motion and accumulation predict ive network (JAMANet) with the guidance of OP-ABA, which is able to efficiently estimate the adversarial motion and accumulation parameters in a one-step way. T he experiments on four popular datasets (e.g., OTB100, VOT2018, UAV123, and LaSO T) demonstrate that our methods are able to cause significant accuracy drops on four state-of-the-art trackers with high transferability. Please find the source code at https://github.com/tsingqguo/ABA.
****
Few-Shot Image Classification: Just Use a Library of Pre-Trained Feature Extract ors and a Simple Classifier
Arkabandhu Chowdhury, Mingchao Jiang, Swarat Chaudhuri, Chris Jermaine; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 9445-9454
Recent papers have suggested that transfer learning can outperform sophisticated meta-learning methods for few-shot image classification. We take this hypothesi s to its logical conclusion, and suggest the use of an ensemble of high-quality, pre-trained feature extractors for few-shot image classification. We show exper imentally that a library of pre-trained feature extractors combined with a simpl e feed-forward network learned with an L2-regularizer can be an excellent option for solving cross-domain few-shot image classification. Our experimental result s suggest that this simpler sample-efficient approach far outperforms several we ll-established meta-learning algorithms.
****
COMISR: Compression-Informed Video Super-Resolution
Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, Peyman Milanfar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2543-2552
Most video super-resolution methods focus on restoring high-resolution video fra mes from low-resolution videos without taking into account compression. However, most videos on the web or mobile devices are compressed, and the compression ca n be severe when the bandwidth is limited. In this paper, we propose a new compr ession-informed video super-resolution model to restore high-resolution content without introducing artifacts caused by compression. The proposed model consists of three modules for video super-resolution: bi-directional recurrent warping, detail-preserving flow estimation, and Laplacian enhancement. All these three mo dules are used to deal with compression properties such as the location of the i ntra-frames in the input and smoothness in the output frames. For thorough perfo rmance evaluation, we conducted extensive experiments on standard datasets with a wide range of compression rates, covering many real video use cases. We showed that our method not only recovers high-resolution content on uncompressed frame

s from the widely-used benchmark datasets, but also achieves state-of-the-art pe rformance in super-resolving compressed videos based on numerous quantitative me trics. We also evaluated the proposed method by simulating streaming from YouTub e to demonstrate its effectiveness and robustness. The source codes and trained models are available at https://github.com/google-research/google-research/tree/ master/comisr.
****

Bit-Mixer: Mixed-Precision Networks With Runtime Bit-Width Selection
Adrian Bulat, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5188-5197
Mixed-precision networks allow for a variable bit-width quantization for every l ayer in the network. A major limitation of existing work is that the bit-width f or each layer must be predefined during training time. This allows little flexib ility if the characteristics of the device on which the network is deployed chan ge during runtime. In this work, we propose Bit-Mixer, the very first method to train a meta-quantized network where during test time any layer can change its b it-width without affecting at all the overall network's ability for highly accur ate inference. To this end, we make 2 key contributions: (a) Transitional Batch-Norms, and (b) a 3-stage optimization process which is shown capable of training such a network. We show that our method can result in mixed precision networks that exhibit the desirable flexibility properties for on-device deployment witho ut compromising accuracy. Code will be made available.
****

Light Field Saliency Detection With Dual Local Graph Learning and Reciprocative Guidance
Nian Liu, Wangbo Zhao, Dingwen Zhang, Junwei Han, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4712-4721
The application of light field data in salient object detection is becoming incr easingly popular in recent years. The difficulty lies in how to effectively fuse the features within the focal stack and how to cooperate them with the feature of the all-focus image. Previous methods usually fuse focal stack features via c onvolution or ConvLSTM, which are both less effective and ill-posed. In this pap er, we model the information fusion within focal stack via graph networks. They introduce powerful context propagation from neighbouring nodes and also avoid il l-posed implementations. On the one hand, we construct local graph connections t hus avoiding prohibitive computational costs of traditional graph networks. On t he other hand, instead of processing the two kinds of data separately, we build a novel dual graph model to guide the focal stack fusion process using all-focus patterns. To handle the second difficulty, previous methods usually implement o ne-shot fusion for focal stack and all-focus features, hence lacking a thorough exploration of their supplements. We introduce a reciprocative guidance scheme a nd enable mutual guidance between these two kinds of information at multiple ste ps. As such, both kinds of features can be enhanced iteratively, finally benefit ing the saliency prediction. Extensive experimental results show that the propos ed models are all beneficial and we achieve significantly better results than st ate-of-the-art methods.
****

Finding Representative Interpretations on Convolutional Neural Networks
Peter Cho-Ho Lam, Lingyang Chu, Maxim Torgonskiy, Jian Pei, Yong Zhang, Lanjun W ang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 1345-1354
Interpreting the decision logic behind effective deep convolutional neural netwo rks (CNN) on images complements the success of deep learning models. However, th e existing methods can only interpret some specific decision logic on individual or a small number of images. To facilitate human understandability and generali zation ability, it is important to develop representative interpretations that i nterpret common decision logics of a CNN on a large group of similar images, whi ch reveal the common semantics data contributes to many closely related predicti ons. In this paper, we develop a novel unsupervised approach to produce a highly representative interpretation for a large number of similar images. We formulat

e the problem of finding representative interpretations as a co-clustering probl
em, and convert it into a submodular cost submodular cover problem based on a sa
mple of the linear decision boundaries of a CNN. We also present a visualization
 and similarity ranking method. Our extensive experiments demonstrate the excell
ent performance of our method.
****

AINet: Association Implantation for Superpixel Segmentation
Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, Yi Yang; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7078-7087
Recently, some approaches are proposed to harness deep convolutional networks to
 facilitate superpixel segmentation. The common practice is to first evenly divi
de the image into a pre-defined number of grids and then learn to associate each
 pixel with its surrounding grids. However, simply applying a series of convolut
ion operations with limited receptive fields can only implicitly perceive the re
lations between the pixel and its surrounding grids. Consequently, existing meth
ods often fail to provide an effective context when inferring the association ma
p. To remedy this issue, we propose a novel Association Implantation (AI) module
 to enable the network to explicitly capture the relations between the pixel and
 its surrounding grids. The proposed AI module directly implants the features of
 grid cells to the surrounding of its corresponding central pixel, and conducts
convolution on the padded window to adaptively transfer knowledge between them.
With such an implantation operation, the network could explicitly harvest the pi
xel-grid level context, which is more in line with the target of superpixel segm
entation comparing to the pixel-wise relation. Furthermore, to pursue better bou
ndary precision, we design a boundary-perceiving loss to help the network discri
minate the pixels around boundaries in hidden feature level, which could benefit
 the subsequent inferring modules to accurately identify more boundary pixels. E
xtensive experiments on BSDS500 and NYUv2 datasets show that our method could no
t only achieve state-of-the-art performance but maintain satisfactory inference
efficiency. Code and pre-trained model are available at https://github.com/wangy
xxjtu/AINet-ICCV2021.
****

An Asynchronous Kalman Filter for Hybrid Event Cameras
Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, Robert Mahony; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 448-457
Event cameras are ideally suited to capture HDR visual information without blur
but perform poorly on static or slowly changing scenes. Conversely, conventional
 image sensors measure absolute intensity of slowly changing scenes effectively
but do poorly on high dynamic range or quickly changing scenes. In this paper, w
e present an event-based video reconstruction pipeline for High Dynamic Range (H
DR) scenarios. The proposed algorithm includes a frame augmentation pre-processi
ng step that deblurs and temporally interpolates frame data using events. The au
gmented frame and event data are then fused using a novel asynchronous Kalman fi
lter under a unifying uncertainty model for both sensors. Our experimental resul
ts are evaluated on both publicly available datasets with challenging lighting c
onditions and fast motions and our new dataset with HDR reference. The proposed
algorithm outperforms state-of-the-art methods in both absolute intensity error
(48% reduction) and image similarity indexes (average 11% improvement).
****

Orthogonal Projection Loss
Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, Fahad Shahbaz
Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2021, pp. 12333-12343
Deep neural networks have achieved remarkable performance on a range of classifi
cation tasks, with softmax cross-entropy (CE) loss emerging as the de-facto obje
ctive function. The CE loss encourages features of a class to have a higher proj
ection score on the true class-vector compared to the negative classes. However,
 this is a relative constraint and does not explicitly force different class fea
tures to be well-separated. Motivated by the observation that ground-truth class
 representations in CE loss are orthogonal (one-hot encoded vectors), we develop

a novel loss function termed `Orthogonal Projection Loss' (OPL) which imposes o
rthogonality in the feature space. OPL augments the properties of CE loss and di
rectly enforces inter-class separation alongside intra-class clustering in the f
eature space through orthogonality constraints on the mini-batch level. As compa
red to other alternatives of CE, OPL offers unique advantages e.g., no additiona
l learnable parameters, does not require careful negative mining and is not sens
itive to the batch size. Given the plug-and-play nature of OPL, we evaluate it o
n a diverse range of tasks including image recognition (CIFAR-100), large-scale
classification (ImageNet), domain generalization (PACS) and few-shot learning (m
iniImageNet, CIFAR-FS, tiered-ImageNet and Meta-dataset) and demonstrate its eff
ectiveness across the board. Furthermore, OPL offers better robustness against p
ractical nuisances such as adversarial attacks and label noise. Our code will be
publicly released.
****

Deep Virtual Markers for Articulated 3D Shapes
Hyomin Kim, Jungeon Kim, Jaewon Kam, Jaesik Park, Seungyong Lee; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11615
-11625
We propose deep virtual markers, a framework for estimating dense and accurate p
ositional information for various types of 3D data. We design a concept and cons
truct a framework that maps 3D points of 3D articulated models, like humans, int
o virtual marker labels. To realize the framework, we adopt a sparse convolution
al neural network and classify 3D points of an articulated model into virtual ma
rker labels. We propose to use soft labels for the classifier to learn rich and
dense interclass relationships based on geodesic distance. To measure the locali
zation accuracy of the virtual markers, we test FAUST challenge, and our result
outperforms the state-of-the-art. We also observe outstanding performance on the
generalizability test, unseen data evaluation, and different 3D data types (mes
hes and depth maps). We show additional applications using the estimated virtual
markers, such as non-rigid registration, texture transfer, and realtime dense m
arker prediction from depth maps.
****

Achieving On-Mobile Real-Time Super-Resolution With Neural Architecture and Prun
ing Search
Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, Wei Niu, Yushu Wu, Tianyun Zhang, Ma
lith Jayaweera, David Kaeli, Bin Ren, Xue Lin, Yanzhi Wang; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4821-4831
Though recent years have witnessed remarkable progress in single image super-res
olution (SISR) tasks with the prosperous development of deep neural networks (DN
Ns), the deep learning methods are confronted with the computation and memory co
nsumption issues in practice, especially for resource-limited platforms such as
mobile devices. To overcome the challenge and facilitate the real-time deploymen
t of SISR tasks on mobile, we combine neural architecture search with pruning se
arch and propose an automatic search framework that derives sparse super-resolut
ion (SR) models with high image quality while satisfying the real-time inference
requirement. To decrease the search cost, we leverage the weight sharing strate
gy by introducing a supernet and decouple the search problem into three stages,
including supernet construction, compiler-aware architecture and pruning search,
and compiler-aware pruning ratio search. With the proposed framework, we are th
e first to achieve real-time SR inference (with only tens of milliseconds per fr
ame) for implementing 720p resolution with competitive image quality (in terms o
f PSNR and SSIM) on mobile platforms (Samsung Galaxy S20).
****

One-Pass Multi-View Clustering for Large-Scale Data
Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Li Liu, Siqi Wang, Weixuan Liang, Jiangy
ong Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 12344-12353
Existing non-negative matrix factorization based multi-view clustering algorithm
s compute multiple coefficient matrices respect to different data views, and lea
rn a common consensus concurrently. The final partition is always obtained from

the consensus with classical clustering techniques, such as k-means. However, th
e non-negativity constraint prevents from obtaining a more discriminative embedd
ing. Meanwhile, this two-step procedure fails to unify multi-view matrix factori
zation with partition generation closely, resulting in unpromising performance.
Therefore, we propose an one-pass multi-view clustering algorithm by removing th
e non-negativity constraint and jointly optimize the aforementioned two steps. I
n this way, the generated partition can guide multi-view matrix factorization to
 produce more purposive coefficient matrix which, as a feedback, improves the qu
ality of partition. To solve the resultant optimization problem, we design an al
ternate strategy which is guaranteed to be convergent theoretically. Moreover, t
he proposed algorithm is free of parameter and of linear complexity, making it p
ractical in applications. In addition, the proposed algorithm is compared with r
ecent advances in literature on benchmarks, demonstrating its effectiveness, sup
eriority and efficiency.
****

Knowledge-Enriched Distributional Model Inversion Attacks
Si Chen, Mostafa Kahla, Ruoxi Jia, Guo-Jun Qi; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 16178-16187
Model inversion (MI) attacks are aimed at reconstructing training data from mode
l parameters. Such attacks have triggered increasing concerns about privacy, esp
ecially given the growing number of online model repositories. However, existing
 MI attacks against deep neural networks (DNNs) have a large room for performanc
e improvement. We present a novel inversion-specific GAN that can better distill
 knowledge useful for performing attacks on private models from public data. In
particular, we train the discriminator to differentiate not only the real and fa
ke samples but the soft-labels provided by the target model. Moreover, unlike pr
evious work that directly searches for a single data point to represent a target
 class, we propose to model a private data distribution for each target class. O
ur experiments show that the combination of these techniques can significantly b
oost the success rate of the state-of-the-art MI attacks by 150%, and generalize
 better to a variety of datasets and models. Our code is available at https://gi
thub.com/SCccc21/Knowledge-Enriched-DMI.
****

Z-Score Normalization, Hubness, and Few-Shot Learning
Nanyi Fei, Yizhao Gao, Zhiwu Lu, Tao Xiang; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2021, pp. 142-151
The goal of few-shot learning (FSL) is to recognize a set of novel classes with
only few labeled samples by exploiting a large set of abundant base class sample
s. Adopting a meta-learning framework, most recent FSL methods meta-learn a deep
 feature embedding network, and during inference classify novel class samples us
ing nearest neighbor in the learned high-dimensional embedding space. This means
 that these methods are prone to the hubness problem, that is, a certain class p
rototype becomes the nearest neighbor of many test instances regardless which cl
asses they belong to. However, this problem is largely ignored in existing FSL s
tudies. In this work, for the first time we show that many FSL methods indeed su
ffer from the hubness problem. To mitigate its negative effects, we further prop
ose to employ z-score feature normalization, a simple yet effective transformati
on, during meta-training. A theoretical analysis is provided on why it helps. Ex
tensive experiments are then conducted to show that with z-score normalization,
the performance of many recent FSL methods can be boosted, resulting in new stat
e-of-the-art on three benchmarks.
****

Dense Interaction Learning for Video-Based Person Re-Identification
Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, Xian-Sheng Hua; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 1490-1501
Video-based person re-identification (re-ID) aims at matching the same person ac
ross video clips. Efficiently exploiting multi-scale fine-grained features while
 building the structural interaction among them is pivotal for its success. In t
his paper, we propose a hybrid framework, Dense Interaction Learning (DenseIL),

that takes the principal advantages of both CNN-based and Attention-based architectures to tackle video-based person re-ID difficulties. DenseIL contains a CNN encoder and a Dense Interaction (DI) decoder. The CNN encoder is responsible for efficiently extracting discriminative spatial features while the DI decoder is designed to densely model spatial-temporal inherent interaction across frames. Different from previous works, we additionally let the DI decoder densely attends to intermediate fine-grained CNN features and that naturally yields multi-grained spatial-temporal representation for each video clip. Moreover, we introduce Spatio-TEmporal Positional Embedding (STEP-Emb) into the DI decoder to investigate the positional relation among the spatial-temporal inputs. Our experiments consistently and significantly outperform all the state-of-the-art methods on multiple standard video-based person re-ID datasets.

****

M3D-VTON: A Monocular-to-3D Virtual Try-On Network

Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, Xiaodan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13239-13249

Virtual 3D try-on can provide an intuitive and realistic view for online shopping and has a huge potential commercial value. However, existing 3D virtual try-on methods mainly rely on annotated 3D human shapes and garment templates, which hinders their applications in practical scenarios. 2D virtual try-on approaches provide a faster alternative to manipulate clothed humans, but lack the rich and realistic 3D representation. In this paper, we propose a novel Monocular-to-3D Virtual Try-On Network (M3D-VTON) that builds on the merits of both 2D and 3D approaches. By integrating 2D information efficiently and learning a mapping that lifts the 2D representation to 3D, we make the first attempt to reconstruct a 3D try-on mesh only taking the target clothing and a person image as inputs. The proposed M3D-VTON includes three modules: 1) The Monocular Prediction Module (MPM) that estimates an initial full-body depth map and accomplishes 2D clothes-person alignment through a novel two-stage warping procedure; 2) The Depth Refinement Module (DRM) that refines the initial body depth to produce more detailed pleat and face characteristics; 3) The Texture Fusion Module (TFM) that fuses the warped clothing with the non-target body part to refine the results. We also construct a high-quality synthesized Monocular-to-3D virtual try-on dataset, in which each person image is associated with a front and a back depth map. Extensive experiments demonstrate that the proposed M3D-VTON can manipulate and reconstruct the 3D human body wearing the given clothing with compelling details and is more efficient than other 3D approaches.

****

Explanations for Occluded Images

Hana Chockler, Daniel Kroening, Youcheng Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1234-1243

Existing algorithms for explaining the output of image classifiers perform poorly on inputs where the object of interest is partially occluded. We present a novel, black-box algorithm for computing explanations that uses a principled approach based on causal theory. We have implemented the method in the DeepCover tool. We obtain explanations that are much more accurate than those generated by the existing explanation tools on images with occlusions and observe a level of performance comparable to the state of the art when explaining images without occlusions.

****

Designing a Practical Degradation Model for Deep Blind Image Super-Resolution

Kai Zhang, Jingyun Liang, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4791-4800

It is widely acknowledged that single image super-resolution (SISR) methods would not perform well if the assumed degradation model deviates from those in real images. Although several degradation models take additional factors into consideration, such as blur, they are still not effective enough to cover the diverse degradations of real images. To address this issue, this paper proposes to design a more complex but practical degradation model that consists of randomly shuffl

ed blur, downsampling and noise degradations. Specifically, the blur is approxim ated by two convolutions with isotropic and anisotropic Gaussian kernels; the do wnsampling is randomly chosen from nearest, bilinear and bicubic interpolations; the noise is synthesized by adding Gaussian noise with different noise levels, adopting JPEG compression with different quality factors, and generating process ed camera sensor noise via reverse-forward camera image signal processing (ISP) pipeline model and RAW image noise model. To verify the effectiveness of the new degradation model, we have trained a deep blind ESRGAN super-resolver and then applied it to super-resolve both synthetic and real images with diverse degradat ions. The experimental results demonstrate that the new degradation model can he lp to significantly improve the practicability of deep super-resolvers, thus pro viding a powerful alternative solution for real SISR applications.
****

Unshuffling Data for Improved Generalization in Visual Question Answering
Damien Teney, Ehsan Abbasnejad, Anton van den Hengel; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 1417-1427
Generalization beyond the training distribution is a core challenge in machine l earning. The common practice of mixing and shuffling examples when training neur al networks may not be optimal in this regard. We show that partitioning the dat a into well-chosen, non-i.i.d. subsets treated as multiple training environments can guide the learning of models with better out-of-distribution generalization . We describe a training procedure to capture the patterns that are stable acros s environments while discarding spurious ones. The method makes a step beyond co rrelation-based learning: the choice of the partitioning allows injecting inform ation about the task that cannot be otherwise recovered from the joint distribut ion of the training data. We demonstrate multiple use cases with the task of vis ual question answering, which is notorious for dataset biases. We obtain signifi cant improvements on VQA-CP, using environments built from prior knowledge, exis ting meta data, or unsupervised clustering. We also get improvements on GQA usin g annotations of "equivalent questions", and on multi-dataset training (VQA v2 / Visual Genome) by treating them as distinct environments.
****

Architecture Disentanglement for Deep Neural Networks
Jie Hu, Liujuan Cao, Tong Tong, Qixiang Ye, Shengchuan Zhang, Ke Li, Feiyue Huan g, Ling Shao, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 672-681
Understanding the inner workings of deep neural networks (DNNs) is essential to provide trustworthy artificial intelligence techniques for practical application s. Existing studies typically involve linking semantic concepts to units or laye rs of DNNs, but fail to explain the inference process. In this paper, we introdu ce neural architecture disentanglement (NAD) to fill the gap. Specifically, NAD learns to disentangle a pre-trained DNN into sub-architectures according to inde pendent tasks, forming information flows that describe the inference processes. We investigate whether, where, and how the disentanglement occurs through experi ments conducted with handcrafted and automatically-searched network architecture s, on both object-based and scene-based datasets. Based on the experimental resu lts, we present three new findings that provide fresh insights into the inner lo gic of DNNs. First, DNNs can be divided into sub-architectures for independent t asks. Second, deeper layers do not always correspond to higher semantics. Third, the connection type in a DNN affects how the information flows across layers, l eading to different disentanglement behaviors. With NAD, we further explain why DNNs sometimes give wrong predictions. Experimental results show that misclassif ied images have a high probability of being assigned to task sub-architectures s imilar to the correct ones. Our code is available at https://github.com/hujiecpp /NAD.
****

Instances As Queries
Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, Wenyu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 6910-6919

We present QueryInst, a new perspective for instance segmentation. QueryInst is a multi-stage end-to-end system that treats instances of interest as learnable queries, enabling query based object detectors, e.g., Sparse R-CNN, to have strong instance segmentation performance. The attributes of instances such as categories, bounding boxes, instance masks, and instance association embeddings are represented by queries in a unified manner. In QueryInst, a query is shared by both detection and segmentation via dynamic convolutions and driven by parallelly-supervised multi-stage learning. We conduct extensive experiments on three challenging benchmarks, i.e., COCO, CityScapes, and YouTube-VIS to evaluate the effectiveness of QueryInst in object detection, instance segmentation, and video instance segmentation tasks. For the first time, we demonstrate that a simple end-to-end query based framework can achieve the state-of-the-art performance in various instance-level recognition tasks. Code is available at https://github.com/hustvl/QueryInst.

****

Omni-GAN: On the Secrets of cGANs and Beyond

Peng Zhou, Lingxi Xie, Bingbing Ni, Cong Geng, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14061-14071

The conditional generative adversarial network (cGAN) is a powerful tool of generating high-quality images, but existing approaches mostly suffer unsatisfying performance or the risk of mode collapse. This paper presents Omni-GAN, a variant of cGAN that reveals the devil in designing a proper discriminator for training the model. The key is to ensure that the discriminator receives strong supervision to perceive the concepts and moderate regularization to avoid collapse. Omni-GAN is easily implemented and freely integrated with off-the-shelf encoding methods (e.g., implicit neural representation, INR). Experiments validate the superior performance of Omni-GAN and Omni-INR-GAN in a wide range of image generation and restoration tasks. In particular, Omni-INR-GAN sets new records on the ImageNet dataset with impressive Inception scores of 262.85 and 343.22 for the image sizes of 128 and 256, respectively, surpassing the previous records by 100+ points. Moreover, leveraging the generator prior, Omni-INR-GAN can extrapolate low-resolution images to arbitrary resolution, even up to x60+ higher resolution. Code is available.

****

ACDC: The Adverse Conditions Dataset With Correspondences for Semantic Driving Scene Understanding

Christos Sakaridis, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10765-10775

Level 5 autonomy for self-driving cars requires a robust visual perception system that can parse input images under any visual condition. However, existing semantic segmentation datasets are either dominated by images captured under normal conditions or are small in scale. To address this, we introduce ACDC, the Adverse Conditions Dataset with Correspondences for training and testing semantic segmentation methods on adverse visual conditions. ACDC consists of a large set of 4006 images which are equally distributed between four common adverse conditions: fog, nighttime, rain, and snow. Each adverse-condition image comes with a high-quality fine pixel-level semantic annotation, a corresponding image of the same scene taken under normal conditions, and a binary mask that distinguishes between intra-image regions of clear and uncertain semantic content. Thus, ACDC supports both standard semantic segmentation and the newly introduced uncertainty-aware semantic segmentation. A detailed empirical study demonstrates the challenges that the adverse domains of ACDC pose to state-of-the-art supervised and unsupervised approaches and indicates the value of our dataset in steering future progress in the field. Our dataset and benchmark are publicly available.

****

Improving De-Raining Generalization via Neural Reorganization

Jie Xiao, Man Zhou, Xueyang Fu, Aiping Liu, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4987-4996

Most existing image de-raining networks could only learn fixed mapping rules between paired rainy/clean images on single synthetic dataset and then stay static

for lifetime. However, since single synthetic dataset merely provides a partial view for the distribution of rain streaks, the deep models well trained on an individual synthetic dataset tend to overfit on this biased distribution. This leads to the inability of these methods to well generalize to complex and changeable real-world rainy scenes, thus limiting their practical applications. In this paper, we try for the first time to accumulate the de-raining knowledge from multiple synthetic datasets on a single network parameter set to improve the de-raining generalization of deep networks. To achieve this goal, we explore Neural Reorganization (NR) to allow the de-raining network to keep a subtle stability-plasticity trade-off rather than naive stabilization after training phase. Specifically, we design our NR algorithm by borrowing the synaptic consolidation mechanism in the biological brain and knowledge distillation. Equipped with our NR algorithm, the deep model can be trained on a list of synthetic rainy datasets by overcoming catastrophic forgetting, making it a general-version de-raining network. Extensive experimental validation shows that due to the successful accumulation of de-raining knowledge, our proposed method can not only process multiple synthetic datasets consistently, but also achieve state-of-the-art results when dealing with real-world rainy images.

****

3D Shape Generation and Completion Through Point-Voxel Diffusion

Linqi Zhou, Yilun Du, Jiajun Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5826-5835

We propose a novel approach for probabilistic generative modeling of 3D shapes. Unlike most existing models that learn to deterministically translate a latent vector to a shape, our model, Point-Voxel Diffusion (PVD), is a unified, probabilistic formulation for unconditional shape generation and conditional, multi-modal shape completion. PVDmarries denoising diffusion models with the hybrid, point-voxel representation of 3D shapes. It can be viewed as a series of denoising steps, reversing the diffusion process from observed point cloud data to Gaussian noise, and is trained by optimizing a variational lower bound to the (conditional) likelihood function. Experiments demonstrate that PVD is capable of synthesizing high-fidelity shapes, completing partial point clouds, and generating multiple completion results from single-view depth scans of real objects.

****

Temporal Knowledge Consistency for Unsupervised Visual Representation Learning

Weixin Feng, Yuanjiang Wang, Lihua Ma, Ye Yuan, Chi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10170-10180

The instance discrimination paradigm has become dominant in unsupervised learning. It always adopts a teacher-student framework, in which the teacher provides embedded knowledge as a supervision signal for the student. The student learns meaningful representations by enforcing instance spatial consistency with the views from the teacher. However, the outputs of the teacher can vary dramatically on the same instance during different training stages, introducing unexpected noise and leading to catastrophic forgetting caused by inconsistent objectives. In this paper, we first integrate instance temporal consistency into current instance discrimination paradigms, and propose a novel and strong algorithm named Temporal Knowledge Consistency (TKC). Specifically, our TKC dynamically ensembles the knowledge of temporal teachers and adaptively selects useful information according to its importance to learning instance temporal consistency. Experimental result shows that TKC can learn better visual representations on both ResNet and AlexNet on linear evaluation protocol while transfer well to downstream tasks. All experiments suggest the good effectiveness and generalization of our method. Code will be made available.

****

Self-Conditioned Probabilistic Learning of Video Rescaling

Yuan Tian, Guo Lu, Xiongkuo Min, Zhaohui Che, Guangtao Zhai, Guodong Guo, Zhiyong Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4490-4499

Bicubic downscaling is a prevalent technique used to reduce the video storage burden or to accelerate the downstream processing speed. However, the inverse upsc

aling step is non-trivial, and the downscaled video may also deteriorate the per formance of downstream tasks. In this paper, we propose a self-conditioned proba bilistic framework for video rescaling to learn the paired downscaling and upsca ling procedures simultaneously. During the training, we decrease the entropy of the information lost in the downscaling by maximizing its probability conditione d on the strong spatial-temporal prior information within the downscaled video. After optimization, the downscaled video by our framework preserves more meaning ful information, which is beneficial for both the upscaling step and the downstr eam tasks, e.g., video action recognition task. We further extend the framework to a lossy video compression system, in which a gradient estimator for non-diffe rential industrial lossy codecs is proposed for the end-to-end training of the w hole system. Extensive experimental results demonstrate the superiority and effe ctiveness of our approach on video rescaling, video compression, and efficient a ction recognition tasks.

****

Unsupervised Image Generation With Infinite Generative Adversarial Networks

Hui Ying, He Wang, Tianjia Shao, Yin Yang, Kun Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14284-14293

Image generation has been heavily investigated in computer vision, where one cor e research challenge is to generate images from arbitrarily complex distribution s with little supervision. Generative Adversarial Networks (GANs) as an implicit approach have achieved great successes in this direction and therefore been emp loyed widely. However, GANs are known to suffer from issues such as mode collaps e, non-structured latent space, being unable to compute likelihoods, etc. In thi s paper, we propose a new unsupervised non-parametric method named mixture of in finite conditional GANs or MIC-GANs, to tackle several GAN issues together, aimi ng for image generation with parsimonious prior knowledge. Through comprehensive evaluations across different datasets, we show that MIC-GANs are effective in s tructuring the latent space and avoiding mode collapse, and outperform state-of- the-art methods. MICGANs are adaptive, versatile, and robust. They offer a promi sing solution to several well-known GAN issues. Code available: github.com/yingh db/MICGANs.

****

SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estima tion

Kai Chen, Qi Dou; Proceedings of the IEEE/CVF International Conference on Comput er Vision (ICCV), 2021, pp. 2773-2782

Category-level 6D object pose estimation aims to predict the position and orient ation for unseen objects, which plays a pillar role in many scenarios such as ro botics and augmented reality. The significant intra-class variation is the bottl eneck challenge in this task yet remains unsolved so far. In this paper, we take advantage of category prior to overcome this problem by innovating a structure- guided prior adaptation scheme to accurately estimate 6D pose for individual obj ects. Different from existing prior-based methods, given one object and its corr esponding category prior, we propose to leverage their structure similarity to d ynamically adapt the prior to the observed object. The prior adaptation intrinsi cally associates the adopted prior with different objects, from which we can acc urately reconstruct the 3D canonical model of the specific object for pose estim ation. To further enhance the structure characteristic of objects, we extract lo w-rank structure points from the dense object point cloud, therefore more effici ently incorporating sparse structural information during prior adaptation. Exten sive experiments on CAMERA25 and REAL275 benchmarks demonstrate significant perf ormance improvement. Project homepage: https://www.cse.cuhk.edu.hk/ kaichen/proj ects/sgpa/sgpa.html.

****

Inferring High-Resolution Traffic Accident Risk Maps Based on Satellite Imagery and GPS Trajectories

Songtao He, Mohammad Amin Sadeghi, Sanjay Chawla, Mohammad Alizadeh, Hari Balakr ishnan, Samuel Madden; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2021, pp. 11977-11985

Traffic accidents cost about 3% of the world's GDP and are the leading cause of death in children and young adults. Accident risk maps are useful tools to monitor and mitigate accident risk. We present a technique to generate high-resolution (5 meters) accident risk maps. At this high resolution, accidents are sparse and risk estimation is limited by bias-variance trade-off. Prior accident risk maps either estimate low-resolution maps that are of low utility (high bias), or they use frequency-based estimation techniques that inaccurately predict where accidents actually happen (high variance). To improve this trade-off, we use an end-to-end deep architecture that can input satellite imagery, GPS trajectories, road maps and the history of accidents. Our evaluation on four metropolitan areas in the US with a total area of 7,488 km2 shows that our technique outperforms prior work in terms of resolution and accuracy.

****

Self-Supervised Product Quantization for Deep Unsupervised Image Retrieval

Young Kyun Jang, Nam Ik Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12085-12094

Supervised deep learning-based hash and vector quantization are enabling fast and large-scale image retrieval systems. By fully exploiting label annotations, they are achieving outstanding retrieval performances compared to the conventional methods. However, it is painstaking to assign labels precisely for a vast amount of training data, and also, the annotation process is error-prone. To tackle these issues, we propose the first deep unsupervised image retrieval method dubbed Self-supervised Product Quantization (SPQ) network, which is label-free and trained in a self-supervised manner. We design a Cross Quantized Contrastive learning strategy that jointly learns codewords and deep visual descriptors by comparing individually transformed images (views). Our method analyzes the image contents to extract descriptive features, allowing us to understand image representations for accurate retrieval. By conducting extensive experiments on benchmarks, we demonstrate that the proposed method yields state-of-the-art results even without supervised pretraining.

****

On Equivariant and Invariant Learning of Object Landmark Representations

Zezhou Cheng, Jong-Chyi Su, Subhransu Maji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9897-9906

Given a collection of images, humans are able to discover landmarks by modeling the shared geometric structure across instances. This idea of geometric equivariance has been widely used for the unsupervised discovery of object landmark representations. In this paper, we develop a simple and effective approach by combining instance-discriminative and spatially-discriminative contrastive learning. We show that when a deep network is trained to be invariant to geometric and photometric transformations, representations emerge from its intermediate layers that are highly predictive of object landmarks. Stacking these across layers in a "hypercolumn" and projecting them using spatially-contrastive learning further improves their performance on matching and few-shot landmark regression tasks. We also present a unified view of existing equivariant and invariant representation learning approaches through the lens of contrastive learning, shedding light on the nature of invariances learned. Experiments on standard benchmarks for landmark learning, as well as a new challenging one we propose, show that the proposed approach surpasses prior state-of-the-art.

****

Rethinking Deep Image Prior for Denoising

Yeonsik Jo, Se Young Chun, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5087-5096

Deep image prior (DIP) serves as a good inductive bias for diverse inverse problems. Among them, denoising is known to be particularly challenging for the DIP due to noise fitting with the requirement of an early stopping. To address the issue, we first analyze the DIP by the notion of effective degrees of freedom (DF) to monitor the optimization progress and propose a principled stopping criterion before fitting to noise without access of a paired ground truth image for Gaussian noise. We also propose the 'stochastic temporal ensemble (STE)' method for

incorporating techniques to further improve DIP's performance for denoising. We additionally extend our method to Poisson noise. Our empirical validations show that given a single noisy image, our method denoises the image while pre- servin g rich textual details. Further, our approach outperforms prior arts in LPIPS by large margins with comparable PSNR and SSIM on seven different datasets.

****

VariTex: Variational Neural Face Textures

Marcel C. Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13890-13899

Deep generative models can synthesize photorealistic images of human faces with novel identities.However, a key challenge to the wide applicability of such techniques is to provide independent control over semantically meaningful parameters: appearance, head pose, face shape, and facial expressions. In this paper, we propose VariTex - to the best of our knowledge the first method that learns a variational latent feature space of neural face textures, which allows sampling of novel identities. We combine this generative model with a parametric face model and gain explicit control over head pose and facial expressions. To generate complete images of human heads, we propose an additive decoder that adds plausible details such as hair. A novel training scheme enforces a pose-independent latent space and in consequence, allows learning a one-to-many mapping between latent codes and pose-conditioned exterior regions. The resulting method can generate geometrically consistent images of novel identities under fine-grained control over head pose, face shape, and facial expressions. This facilitates a broad range of downstream tasks, like sampling novel identities, changing the head pose, expression transfer, and more.

****

Domain Adaptive Semantic Segmentation With Self-Supervised Depth Estimation

Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, Olga Fink; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8515-8525

Domain adaptation for semantic segmentation aims to improve the model performance in the presence of a distribution shift between source and target domain. Leveraging the supervision from auxiliary tasks (such as depth estimation) has the potential to heal this shift because many visual tasks are closely related to each other. However, such a supervision is not always available. In this work, we leverage the guidance from self-supervised depth estimation, which is available on both domains, to bridge the domain gap. On the one hand, we propose to explicitly learn the task feature correlation to strengthen the target semantic predictions with the help of target depth estimation. On the other hand, we use the depth prediction discrepancy from source and target depth decoders to approximate the pixel-wise adaptation difficulty. The adaptation difficulty, inferred from depth, is then used to refine the target semantic segmentation pseudo-labels. The proposed method can be easily implemented into existing segmentation frameworks. We demonstrate the effectiveness of our approach on the benchmark tasks SYNTHIA-to-Cityscapes and GTA-to-Cityscapes, on which we achieve the new state-of-the-art performance of 55.0% and 56.6%, respectively. Our code is available at https://qin.ee/corda

****

The Way to My Heart Is Through Contrastive Learning: Remote Photoplethysmography From Unlabelled Video

John Gideon, Simon Stent; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3995-4004

The ability to reliably estimate physiological signals from video is a powerful tool in low-cost, pre-clinical health monitoring. In this work we propose a new approach to remote photoplethysmography (rPPG) -- the measurement of blood volume changes from observations of a person's face or skin. Similar to current state-of-the-art methods for rPPG, we apply neural networks to learn deep representations with invariance to nuisance image variation. In contrast to such methods, we employ a fully self-supervised training approach, which has no reliance on expensive ground truth physiological training data. Our proposed method uses contra

stive learning with a weak prior over the frequency and temporal smoothness of t
he target signal of interest. We evaluate our approach on four rPPG datasets, sh
owing that comparable or better results can be achieved compared to recent super
vised deep learning methods but without using any annotation. In addition, we in
corporate a learned saliency resampling module into both our unsupervised approa
ch and supervised baseline. We show that by allowing the model to learn where to
 sample the input image, we can reduce the need for hand-engineered features whi
le providing some interpretability into the model's behavior and possible failur
e modes. We release code for our complete training and evaluation pipeline to en
courage reproducible progress in this exciting new direction.
****

IICNet: A Generic Framework for Reversible Image Conversion
Ka Leong Cheng, Yueqi Xie, Qifeng Chen; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2021, pp. 1991-2000
Reversible image conversion (RIC) aims to build a reversible transformation betw
een specific visual content (e.g., short videos) and an embedding image, where t
he original content can be restored from the embedding when necessary. This work
 develops Invertible Image Conversion Net (IICNet) as a generic solution to vari
ous RIC tasks due to its strong capacity and task-independent design. Unlike pre
vious encoder-decoder based methods, IICNet maintains a highly invertible struct
ure based on invertible neural networks (INNs) to better preserve the informatio
n during conversion. We use a relation module and a channel squeeze layer to imp
rove the INN nonlinearity to extract cross-image relations and the network flexi
bility, respectively. Experimental results demonstrate that IICNet outperforms t
he specifically-designed methods on existing RIC tasks and can generalize well t
o various newly-explored tasks. With our generic IICNet, we no longer need to ha
nd-engineer task-specific embedding networks for rapidly occurring visual conten
t. Our source codes are available at: https://github.com/felixcheng97/IICNet.
****

Deep Hough Voting for Robust Global Registration
Junha Lee, Seungwook Kim, Minsu Cho, Jaesik Park; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2021, pp. 15994-16003
Point cloud registration is the task of estimating the rigid transformation that
 aligns a pair of point cloud fragments. We present an efficient and robust fram
ework for pairwise registration of real-world 3D scans, leveraging Hough voting
in the 6D transformation parameter space. First, deep geometric features are ext
racted from a point cloud pair to compute putative correspondences. We then cons
truct a set of triplets of correspondences to cast votes on the 6D Hough space,
which represents the transformation parameters in the form of sparse tensors. Ne
xt, a fully convolutional refinement module is applied to refine the noisy votes
. Finally, we identify the consensus among the correspondences from the Hough sp
ace, which we use to predict our final transformation parameters. Our method out
performs state-of-the-art methods on the 3DMatch and 3DLoMatch benchmarks while
achieving comparable performance on the KITTI odometry dataset. We further demon
strate the generalizability of our approach by setting a new state-of-the-art on
 the ICL-NUIM dataset, where we integrate our module into a multi-way registrati
on pipeline.
****

Image Synthesis From Layout With Locality-Aware Mask Adaption
Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, Lingyun Sun; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13819
-13828
This paper is concerned with synthesizing images conditioned on a layout (a set
of bounding boxes with object categories). Existing works construct a layout-mas
k-image pipeline. Object masks are generated separately and mapped to bounding b
oxes to form a whole semantic segmentation mask (layout-to-mask), with which a n
ew image is generated (mask-to-image). However, overlapped boxes in layouts resu
lt in overlapped object masks, which reduces the mask clarity and causes confusi
on in image generation. We hypothesize the importance of generating clean and se
mantically clear semantic masks. The hypothesis is supported by the finding that

the performance of state-of-the-art LostGAN decreases when input masks are tainted. Motivated by this hypothesis, we propose Locality-Aware Mask Adaption (LAMA) module to adapt overlapped or nearby object masks in the generation. Experimental results show our proposed model with LAMA outperforms existing approaches regarding visual fidelity and alignment with input layouts. On COCO-stuff in 256x256, our method improves the state-of-the-art FID score from 41.65 to 31.12 and the SceneFID from 22.00 to 18.64.
****

Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration Without Forgetting
Anna Kukleva, Hilde Kuehne, Bernt Schiele; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9020-9029
Both generalized and incremental few-shot learning have to deal with three major challenges: learning novel classes from only few samples per class, preventing catastrophic forgetting of base classes, and classifier calibration across novel and base classes. In this work we propose a three-stage framework that allows to explicitly and effectively address these challenges. While the first phase learns base classes with many samples, the second phase learns a calibrated classifier for novel classes from few samples while also preventing catastrophic forgetting. In the final phase, calibration is achieved across all classes. We evaluate the proposed framework on four challenging benchmark datasets for image and video few-shot classification and obtain state-of-the-art results for both generalized and incremental few shot learning.
****

Scribble-Supervised Semantic Segmentation by Uncertainty Reduction on Neural Representation and Self-Supervision on Neural Eigenspace
Zhiyi Pan, Peng Jiang, Yunhai Wang, Changhe Tu, Anthony G. Cohn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7416-7425
Scribble-supervised semantic segmentation has gained much attention recently for its promising performance without high-quality annotations. Due to the lack of supervision, confident and consistent predictions are usually hard to obtain. Typically, people handle these problems by either adopting an auxiliary task with the well-labeled dataset or incorporating a graphical model with additional requirements on scribble annotations. Instead, this work aims to achieve semantic segmentation by scribble annotations directly without extra information and other limitations. Specifically, we propose holistic operations, including minimizing entropy and a network embedded random walk on the neural representation to reduce uncertainty. Given the probabilistic transition matrix of a random walk, we further train the network with self-supervision on its neural eigenspace to impose consistency on predictions between related images. Comprehensive experiments and ablation studies verify the proposed approach, which demonstrates superiority over others; it is even comparable to some full-label supervised ones and works well when scribbles are randomly shrunk or dropped.
****

Unsupervised Domain Adaptive 3D Detection With Multi-Level Consistency
Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8866-8875
Deep learning-based 3D object detection has achieved unprecedented success with the advent of large-scale autonomous driving datasets. However, drastic performance degradation remains a critical challenge for cross-domain deployment. In addition, existing 3D domain adaptive detection methods often assume prior access to the target domain annotations, which is rarely feasible in the real world. To address this challenge, we study a more realistic setting, unsupervised 3D domain adaptive detection, which only utilizes source domain annotations. 1) We first comprehensively investigate the major underlying factors of the domain gap in 3D detection. Our key insight is that geometric mismatch is the key factor of domain shift. 2) Then, we propose a novel and unified framework, Multi-Level Consistency Network (MLC-Net), which employs a teacher-student paradigm to generate ad

aptive and reliable pseudo-targets. MLC-Net exploits point-, instance- and neural statistics-level consistency to facilitate cross-domain transfer. Extensive experiments demonstrate that MLC-Net outperforms existing state-of-the-art methods (including those using additional target domain information) on standard benchmarks. Notably, our approach is detector-agnostic, which achieves consistent gains on both single- and two-stage 3D detectors. Code will be released.
****

Transporting Causal Mechanisms for Unsupervised Domain Adaptation
Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, Hanwang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8599-8608
Existing Unsupervised Domain Adaptation (UDA) literature adopts the covariate shift and conditional shift assumptions, which essentially encourage models to learn common features across domains. However, due to the lack of supervision in the target domain, they suffer from the semantic loss: the feature will inevitably lose non-discriminative semantics in source domain, which is however discriminative in target domain. We use a causal view---transportability theory---to identify that such loss is in fact a confounding effect, which can only be removed by causal intervention. However, the theoretical solution provided by transportability is far from practical for UDA, because it requires the stratification and representation of the unobserved confounder that is the cause of the domain gap. To this end, we propose a practical solution: Transporting Causal Mechanisms (TCM), to identify the confounder stratum and representations by using the domain-invariant disentangled causal mechanisms, which are discovered in an unsupervised fashion. Our TCM is both theoretically and empirically grounded. Extensive experiments show that TCM achieves state-of-the-art performance on three challenging UDA benchmarks: ImageCLEF-DA, Office-Home, and VisDA-2017. Codes are available at https://github.com/yue-zhongqi/tcm.
****

Learning To Estimate Hidden Motions With Global Motion Aggregation
Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, Richard Hartley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9772-9781
Occlusions pose a significant challenge to optical flow algorithms that rely on local evidences. We consider an occluded point to be one that is imaged in the first frame but not in the next, a slight overloading of the standard definition since it also includes points that move out-of-frame. Estimating the motion of these points is extremely difficult, particularly in the two-frame setting. Previous work relies on CNNs to learn occlusions, without much success, or requires multiple frames to reason about occlusions using temporal smoothness. In this paper, we argue that the occlusion problem can be better solved in the two-frame case by modelling image self-similarities. We introduce a global motion aggregation module, a transformer-based approach to find long-range dependencies between pixels in the first image, and perform global aggregation on the corresponding motion features. We demonstrate that the optical flow estimates in the occluded regions can be significantly improved without damaging the performance in non-occluded regions. This approach obtains new state-of-the-art results on the challenging Sintel dataset, improving the average end-point error by 13.6% on Sintel Final and 13.7% on Sintel Clean. At the time of submission, our method ranks first on these benchmarks among all published and unpublished approaches. Code is available at https://github.com/zacjiang/GMA.
****

Predicting With Confidence on Unseen Distributions
Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, Ludwig Schmidt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1134-1144
Recent work has shown that the accuracy of machine learning models can vary substantially when evaluated on a distribution that even slightly differs from that of the training data. As a result, predicting model performance on previously unseen distributions without access to labeled data is an important challenge with implications for increasing the reliability of machine learning models. In the

context of distribution shift, distance measures are often used to adapt models and improve their performance on new domains, however accuracy estimation is seldom explored in these investigations. Our investigation determines that common distributional distances such as Frechet distance or Maximum Mean Discrepancy, fail to induce reliable estimates of performance under distribution shift. On the other hand, we find that our proposed difference of confidences (DoC) approach yields successful estimates of a classifier's performance over a variety of shifts and model architectures. Despite its simplicity, we observe that DoC outperforms other methods across synthetic, natural, and adversarial distribution shifts, reducing error by (>46%) on several realistic and challenging datasets such as ImageNet-Vid-Robust and ImageNet-Rendition.

****

TAM: Temporal Adaptive Module for Video Recognition

Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, Tong Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13708-13718

Video data is with complex temporal dynamics due to various factors such as camera motion, speed variation, and different activities. To effectively capture this diverse motion pattern, this paper presents a new temporal adaptive module (TAM) to generate video-specific temporal kernels based on its own feature map. TAM proposes a unique two-level adaptive modeling scheme by decoupling the dynamic kernel into a location sensitive importance map and a location invariant aggregation weight. The importance map is learned in a local temporal window to capture short-term information, while the aggregation weight is generated from a global view with a focus on long-term structure. TAM is a modular block and could be integrated into 2D CNNs to yield a powerful video architecture (TANet) with a very small extra computational cost. The extensive experiments on Kinetics-400 and Something-Something datasets demonstrate that our TAM outperforms other temporal modeling methods consistently, and achieves the state-of-the-art performance under the similar complexity. The code is available at https://github.com/liu-zhy/temporal-adaptive-module.

****

Generating Masks From Boxes by Mining Spatio-Temporal Consistencies in Videos

Bin Zhao, Goutam Bhat, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13556-13566

Segmenting objects in videos is a fundamental computer vision task. The current deep learning based paradigm offers a powerful, but data-hungry solution. However, current datasets are limited by the cost and human effort of annotating object masks in videos. This effectively limits the performance and generalization capabilities of existing video segmentation methods. To address this issue, we explore weaker form of bounding box annotations. We introduce a method for generating segmentation masks from per-frame bounding box annotations in videos. To this end, we propose a spatio-temporal aggregation module that effectively mines consistencies in the object and background appearance across multiple frames. We use our predicted accurate masks to train video object segmentation (VOS) networks for the tracking domain, where only manual bounding box annotations are available. The additional data provides substantially better generalization performance, leading to state-of-the-art results on standard tracking benchmarks. The code and models are available at https://github.com/visionml/pytracking.

****

TRAR: Routing the Attention Spans in Transformer for Visual Question Answering

Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2074-2084

Due to the superior ability of global dependency modeling, Transformer and its variants have become the primary choice of many vision-and-language tasks. However, in tasks like Visual Question Answering (VQA) and Referring Expression Comprehension (REC), the multimodal prediction often requires visual information from macro- to micro-views. Therefore, how to dynamically schedule the global and local dependency modeling in Transformer has become an emerging issue. In this pape

r, we propose an example-dependent routing scheme called TRAnsformer Routing (TR
AR) to address this issue. Specifically, in TRAR, each visual Transformer layer
is equipped with a routing module with different attention spans. The model can
dynamically select the corresponding attentions based on the output of the previ
ous inference step, so as to formulate the optimal routing path for each example
. Notably, with careful designs, TRAR can reduce the additional computation and
memory overhead to almost negligible. To validate TRAR, we conduct extensive exp
eriments on five benchmark datasets of VQA and REC, and achieve superior perform
ance gains than the standard Transformers and a bunch of state-of-the-art method
s.
****

Embed Me if You Can: A Geometric Perceptron
Pavlo Melnyk, Michael Felsberg, Mårten Wadenbäck; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2021, pp. 1276-1284
Solving geometric tasks involving point clouds by using machine learning is a ch
allenging problem. Standard feed-forward neural networks combine linear or, if t
he bias parameter is included, affine layers and activation functions. Their geo
metric modeling is limited, which motivated the prior work introducing the multi
layer hypersphere perceptron (MLHP). Its constituent part, i.e., the hypersphere
 neuron, is obtained by applying a conformal embedding of Euclidean space. By vi
rtue of Clifford algebra, it can be implemented as the Cartesian dot product of
inputs and weights. If the embedding is applied in a manner consistent with the
dimensionality of the input space geometry, the decision surfaces of the model u
nits become combinations of hyperspheres and make the decision-making process ge
ometrically interpretable for humans. Our extension of the MLHP model, the multi
layer geometric perceptron (MLGP), and its respective layer units, i.e., geometr
ic neurons, are consistent with the 3D geometry and provide a geometric handle o
f the learned coefficients. In particular, the geometric neuron activations are
isometric in 3D, which is necessary for rotation and translation equivariance. W
hen classifying the 3D Tetris shapes, we quantitatively show that our model requ
ires no activation function in the hidden layers other than the embedding to out
perform the vanilla multilayer perceptron. In the presence of noise in the data,
 our model is also superior to the MLHP.
****

Learning Rare Category Classifiers on a Tight Labeling Budget
Ravi Teja Mullapudi, Fait Poms, William R. Mark, Deva Ramanan, Kayvon Fatahalian
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2021, pp. 8423-8432
Many real-world ML deployments face the challenge of training a rare category mo
del with a small labeling bud- get. In these settings, there is often access to
large amounts of unlabeled data, therefore it is attractive to consider semi-sup
ervised or active learning approaches to reduce human labeling effort. However,
prior approaches make two assumptions that do not often hold in practice; (a) on
e has access to a modest amount of labeled data to bootstrap learning and (b) ev
ery image belongs to a common category of interest. In this paper, we consider t
he scenario where we start with as-little-as five labeled positives of a rare ca
tegory and a large amount of unlabeled data of which 99.9% of it is negatives. W
e propose an active semi-supervised method for building accurate models in this
challenging setting. Our method leverages two key ideas: (a) Utilize human and m
achine effort where they are most effective; human labels are used to identify "
needle-in-a-haystack" positives, while machine-generated pseudo-labels are used
to identify negatives. (b) Adapt recently proposed representation learning techn
iques for handling extremely imbalanced human labeled data to iteratively train
models with noisy machine labeled data. We compare our approach with prior activ
e learning and semi-supervised approaches, demonstrating significant improvement
s in accuracy per unit labeling effort, particularly on a tight labeling budget.
****

Persistent Homology Based Graph Convolution Network for Fine-Grained 3D Shape Se
gmentation
Chi-Chong Wong, Chi-Man Vong; Proceedings of the IEEE/CVF International Conferen

ce on Computer Vision (ICCV), 2021, pp. 7098-7107
Fine-grained 3D segmentation is an important task in 3D object understanding, especially in applications such as intelligent manufacturing or parts analysis for 3D objects. However, many challenges involved in such problem are yet to be solved, such as i) interpreting the complex structures located in different regions for 3D objects; ii) capturing fine-grained structures with sufficient topology correctness. Current deep learning and graph machine learning methods fail to tackle such challenges and thus provide inferior performance in fine-grained 3D analysis. In this work, methods in topological data analysis are incorporated with geometric deep learning model for the task of fine-grained segmentation for 3D objects. We propose a novel neural network model called Persistent Homology based Graph Convolution Network (PHGCN), which i) integrates persistent homology into graph convolution network to capture multi-scale structural information that can accurately represent complex structures for 3D objects; ii) applies a novel Persistence Diagram Loss that provides sufficient topology correctness for segmentation over the fine-grained structures. Extensive experiments on fine-grained 3D segmentation validate the effectiveness of the proposed PHGCN model and show significant improvements over current state-of-the-art methods.
****
Hybrid Neural Fusion for Full-Frame Video Stabilization
Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, Jia-Bin Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2299-2308
Existing video stabilization methods often generate visible distortion or require aggressive cropping of frame boundaries, resulting in smaller field of views. In this work, we present a frame synthesis algorithm to achieve full-frame video stabilization. We first estimate dense warp fields from neighboring frames and then synthesize the stabilized frame by fusing the warped contents. Our core technical novelty lies in the learning-based hybrid-space fusion that alleviates artifacts caused by optical flow inaccuracy and fast-moving objects. We validate the effectiveness of our method on the NUS, selfie, and DeepStab video datasets. Extensive experiment results demonstrate the merits of our approach over prior video stabilization methods.
****
HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training With Crafted Input Noise
Souvik Kundu, Massoud Pedram, Peter A. Beerel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5209-5218
Low-latency deep spiking neural networks (SNNs) have become a promising alternative to conventional artificial neural networks (ANNs) because of their potential for increased energy efficiency on event-driven neuromorphic hardware. Neural networks, including SNNs, however, are subject to various adversarial attacks and must be trained to remain resilient against such attacks for many applications. Nevertheless, due to prohibitively high training costs associated with SNNs, analysis, and optimization of deep SNNs under various adversarial attacks have been largely overlooked. In this paper, we first present a detailed analysis of the inherent robustness of low-latency SNNs against popular gradient-based attacks, namely fast gradient sign method (FGSM) and projected gradient descent (PGD). Motivated by this analysis, to harness the model robustness against these attacks we present an SNN training algorithm that uses crafted input noise and incurs no additional training time. To evaluate the merits of our algorithm, we conducted extensive experiments with variants of VGG and ResNet on both CIFAR-10 and CIFAR-100 datasets. Compared to standard trained direct input SNNs, our trained models yield improved classification accuracy of up to 13.7% and 10.1% on FGSM and PGD attack-generated images, respectively, with negligible loss in clean image accuracy. Our models also outperform inherently-robust SNNs trained on rate-coded inputs with improved or similar classification performance on attack-generated im-ages while having up to 25x and 4.6x lower latency and computation energy, respectively.
****

CDNet: Centripetal Direction Network for Nuclear Instance Segmentation
Hongliang He, Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Qian Ren, Pengxu We i, Zhiqiang Gao, Jie Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4026-4035
Nuclear instance segmentation is a challenging task due to a large number of tou ching and overlapping nuclei in pathological images. Existing methods cannot eff ectively recognize the accurate boundary owing to neglecting the relationship be tween pixels (e.g., direction information). In this paper, we propose a novel Ce ntripetal Direction Network (CDNet) for nuclear instance segmentation. Specifica lly, we define the centripetal direction feature as a class of adjacent directio ns pointing to the nuclear center to represent the spatial relationship between pixels within the nucleus. These direction features are then used to construct a direction difference map to represent the similarity within instances and the d ifferences between instances. Finally, we propose a direction-guided refinement module, which acts as a plug-and-play module to effectively integrate auxiliary tasks and aggregate the features of different branches. Experiments on MoNuSeg a nd CPM17 datasets show that CDNet is significantly better than the other methods and achieves the state-of-the-art performance. The code is available at https:/ /github.com/honglianghe/CDNet.
****

3D Human Texture Estimation From a Single Image With Transformers
Xiangyu Xu, Chen Change Loy; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2021, pp. 13849-13858
We propose a Transformer-based framework for 3D human texture estimation from a single image. The proposed Transformer is able to effectively exploit the global information of the input image, overcoming the limitations of existing methods that are solely based on convolutional neural networks. In addition, we also pro pose a mask-fusion strategy to combine the advantages of the RGB-based and textu re-flow-based models. We further introduce a part-style loss to help reconstruct high-fidelity colors without introducing unpleasant artifacts. Extensive experi ments demonstrate the effectiveness of the proposed method against state-of-the- art 3D human texture estimation approaches both quantitatively and qualitatively .
****

The Surprising Effectiveness of Visual Odometry Techniques for Embodied PointGoa l Navigation
Xiaoming Zhao, Harsh Agrawal, Dhruv Batra, Alexander G. Schwing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16127 -16136
It is fundamental for personal robots to reliably navigate to a specified goal. To study this task, PointGoal navigation has been introduced in simulated Embodi ed AI environments. Recent advances solve this PointGoal navigation task with ne ar-perfect accuracy (99.6% success) in photo-realistically simulated environment s, assuming noiseless egocentric vision, noiseless actuation, and most important ly, perfect localization. However, under realistic noise models for visual senso rs and actuation, and without access to a "GPS and Compass sensor," the 99.6%-su ccess agents for PointGoal navigation only succeed with 0.3%. In this work, we d emonstrate the surprising effectiveness of visual odometry for the task of Point Goal navigation in this realistic setting, i.e., with realistic noise models for perception and actuation and without access to GPS and Compass sensors. We show that integrating visual odometry techniques into navigation policies improves t he state-of-the-art on the popular Habitat PointNav benchmark by a large margin, improving success from 64.5% to 71.7% while executing 6.4 times faster.
****

Rehearsal Revealed: The Limits and Merits of Revisiting Samples in Continual Lea rning
Eli Verwimp, Matthias De Lange, Tinne Tuytelaars; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2021, pp. 9385-9394
Learning from non-stationary data streams and overcoming catastrophic forgetting still poses a serious challenge for machine learning research. Rather than aimi

ng to improve state-of-the-art, in this work we provide insight into the limits and merits of rehearsal, one of continual learning's most established methods. We hypothesize that models trained sequentially with rehearsal tend to stay in the same low-loss region after a task has finished, but are at risk of overfitting on its sample memory, hence harming generalization. We provide both conceptual and strong empirical evidence on three benchmarks for both behaviors, bringing novel insights into the dynamics of rehearsal and continual learning in general. Finally, we interpret important continual learning works in the light of our findings, allowing for a deeper understanding of their successes.
****

Group-Free 3D Object Detection via Transformers
Ze Liu, Zheng Zhang, Yue Cao, Han Hu, Xin Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2949-2958
Recently, directly detecting 3D objects from 3D point clouds has received increasing attention. To extract object representation from an irregular point cloud, existing methods usually take a point grouping step to assign the points to an object candidate so that a PointNet-like network could be used to derive object features from the grouped points. However, the inaccurate point assignments caused by the hand-crafted grouping scheme decrease the performance of 3D object detection. In this paper, we present a simple yet effective method for directly detecting 3D objects from the 3D point cloud. Instead of grouping local points to each object candidate, our method computes the feature of an object from all the points in the point cloud with the help of an attention mechanism in the Transformers, where the contribution of each point is automatically learned in the network training. With an improved attention stacking scheme, our method fuses object features in different stages and generates more accurate object detection results. With few bells and whistles, the proposed method achieves state-of-the-art 3D object detection performance on two widely used benchmarks, ScanNet V2 and SUN RGB-D.
****

Discover the Unknown Biased Attribute of an Image Classifier
Zhiheng Li, Chenliang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14970-14979
Recent works find that AI algorithms learn biases from data. Therefore, it is urgent and vital to identify biases in AI algorithms. However, the previous bias identification pipeline overly relies on human experts to conjecture potential biases (e.g., gender), which may neglect other underlying biases not realized by humans. To help human experts better find the AI algorithms' biases, we study a new problem in this work -- for a classifier that predicts a target attribute of the input image, discover its unknown biased attribute. To solve this challenging problem, we use a hyperplane in the generative model's latent space to represent an image attribute; thus, the original problem is transformed to optimizing the hyperplane's normal vector and offset. We propose a novel total-variation loss within this framework as the objective function and a new orthogonalization penalty as a constraint. The latter prevents trivial solutions in which the discovered biased attribute is identical with the target or one of the known-biased attributes. Extensive experiments on both disentanglement datasets and real-world datasets show that our method can discover biased attributes and achieve better disentanglement w.r.t. target attributes. Furthermore, the qualitative results show that our method can discover unnoticeable biased attributes for various object and scene classifiers, proving our method's generalizability for detecting biased attributes in diverse domains of images.
****

Learn To Cluster Faces via Pairwise Classification
Junfu Liu, Di Qiu, Pengfei Yan, Xiaolin Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3845-3853
Face clustering plays an essential role in exploiting massive unlabeled face data. Recently, graph-based face clustering methods are getting popular for their satisfying performances. However, they usually suffer from excessive memory consumption especially on large-scale graphs, and rely on empirical thresholds to det

ermine the connectivities between samples in inference, which restricts their ap
plications in various real-world scenes. To address such problems, in this paper
, we explore face clustering from the pairwise angle. Specifically, we formulate
 the face clustering task as a pairwise relationship classification task, avoidi
ng the memory-consuming learning on large-scale graphs. The classifier can direc
tly determine the relationship between samples and is enhanced by taking advanta
ge of the contextual information. Moreover, to further facilitate the efficiency
 of our method, we propose a rank-weighted density to guide the selection of pai
rs sent to the classifier. Experimental results demonstrate that our method achi
eves state-of-the-art performances on several public clustering benchmarks at th
e fastest speed and shows a great advantage in comparison with graph-based clust
ering methods on memory consumption.
****

DAE-GAN: Dynamic Aspect-Aware GAN for Text-to-Image Synthesis
Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, Enhong Chen; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
21, pp. 13960-13969
Text-to-image synthesis refers to generating an image from a given text descript
ion, the key goal of which lies in photo realism and semantic consistency. Previ
ous methods usually generate an initial image with sentence embedding and then r
efine it with fine-grained word embedding. Despite the significant progress, the
 'aspect' information (e.g., red eyes) contained in the text, referring to sever
al words rather than a word that depicts 'a particular part or feature of someth
ing', is often ignored, which is highly helpful for synthesizing image details.
How to make better utilization of aspect information in text-to-image synthesis
still remains an unresolved challenge. To address this problem, in this paper, w
e propose a Dynamic Aspect-awarE GAN (DAE-GAN) that represents text information
comprehensively from multiple granularities, including sentence-level, word-leve
l, and aspect-level. Moreover, inspired by human learning behaviors, we develop
a novel Aspect-aware Dynamic Re-drawer (ADR) for image refinement, in which an A
ttended Global Refinement (AGR) module and an Aspect-aware Local Refinement (ALR
) module are alternately employed. AGR utilizes word-level embedding to globally
 enhance the previously generated image, while ALR dynamically employs aspect-le
vel embedding to refine image details from a local perspective. Finally, a corre
sponding matching loss function is designed to ensure the text-image semantic co
nsistency at different levels. Extensive experiments on two well-studied and pub
licly available datasets (i.e., CUB-200 and COCO) demonstrate the superiority an
d rationality of our method.
****

Learning Facial Representations From the Cycle-Consistency of Face
Jia-Ren Chang, Yong-Sheng Chen, Wei-Chen Chiu; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 9680-9689
Faces manifest large variations in many aspects, such as identity, expression, p
ose, and face styling. Therefore, it is a great challenge to disentangle and ext
ract these characteristics from facial images, especially in an unsupervised man
ner. In this work, we introduce cycle-consistency in facial characteristics as f
ree supervisory signal to learn facial representations from unlabeled facial ima
ges. The learning is realized by superimposing the facial motion cycle-consisten
cy and identity cycle-consistency constraints. The main idea of the facial motio
n cycle-consistency is that, given a face with expression, we can perform de-exp
ression to a neutral face via the removal of facial motion and further perform r
e-expression to reconstruct back to the original face. The main idea of the iden
tity cycle-consistency is to exploit both de-identity into mean face by deprivin
g the given neutral face of its identity via feature re-normalization and re-ide
ntity into neutral face by adding the personal attributes to the mean face. At t
raining time, our model learns to disentangle two distinct facial representation
s to be useful for performing cycle-consistent face reconstruction. At test time
, we use the linear protocol scheme for evaluating facial representations on var
ious tasks, including facial expression recognition and head pose regression. We
 also can directly apply the learnt facial representations to person recognition

, frontalization and image-to-image translation. Our experiments show that the r
esults of our approach is competitive with those of existing methods, demonstrat
ing the rich and unique information embedded in the disentangled representations
. Code is available at https://github.com/JiaRenChang/FaceCycle.
****

Towards Memory-Efficient Neural Networks via Multi-Level In Situ Generation
Jiaqi Gu, Hanqing Zhu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, Da
vid Z. Pan; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2021, pp. 5229-5238
Deep neural networks (DNN) have shown superior performance in a variety of tasks
. As they rapidly evolve, their escalating computation and memory demands make i
t challenging to deploy them on resource-constrained edge devices. Though extens
ive efficient accelerator designs, from traditional electronics to emerging phot
onics, have been successfully demonstrated, they are still bottlenecked by expen
sive memory accesses due to tremendous gaps between the bandwidth/power/latency
of electrical memory and computing cores. Previous solutions fail to fully-lever
age the ultra-fast computational speed of emerging DNN accelerators to break thr
ough the critical memory bound. In this work, we propose a general and unified f
ramework to trade expensive memory transactions with ultra-fast on-chip computat
ions, directly translating to performance improvement. We are the first to joint
ly explore the intrinsic correlations and bit-level redundancy within DNN kernel
s and propose a multi-level in situ generation mechanism with mixed-precision ba
ses to achieve on-the-fly recovery of high-resolution parameters with minimum ha
rdware overhead. Extensive experiments demonstrate that our proposed joint metho
d can boost the memory efficiency by 10-20x with comparable accuracy over four s
tate-of-the-art designs when benchmarked on ResNet-18/DenseNet-121/MobileNetV2/V
3 with various tasks.
****

Greedy Gradient Ensemble for Robust Visual Question Answering
Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, Qi Tian; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1584-1593
Language bias is a critical issue in Visual Question Answering (VQA), where mode
ls often exploit dataset biases for the final decision without considering the i
mage information. As a result, they suffer from performance drop on out-of-distr
ibution data and inadequate visual explanation. Based on experimental analysis f
or existing robust VQA methods, we stress the language bias in VQA that comes fr
om two aspects, i.e., distribution bias and shortcut bias. We further propose a
new de-bias framework, Greedy Gradient Ensemble (GGE), which combines multiple b
iased models for unbiased base model learning. With the greedy strategy, GGE for
ces the biased models to over-fit the biased data distribution in priority, thus
 makes the base model pay more attention to examples that are hard to solve by b
iased models. The experiments demonstrate that our method makes better use of v
isual information and achieves state-of-the-art performance on diagnosing datase
t VQA-CP without using extra annotations.
****

Influence Selection for Active Learning
Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, Conghui He; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 9274-9283
The existing active learning methods select the samples by evaluating the sample
's uncertainty or its effect on the diversity of labeled datasets based on diffe
rent task-specific or model-specific criteria. In this paper, we propose the Inf
luence Selection for Active Learning(ISAL) which selects the unlabeled samples t
hat can provide the most positive Influence on model performance. To obtain the
Influence of the unlabeled sample in the active learning scenario, we design the
 Untrained Unlabeled sample Influence Calculation(UUIC) to estimate the unlabele
d sample's expected gradient with which we calculate its Influence. To prove the
 effectiveness of UUIC, we provide both theoretical and experimental analyses. S
ince the UUIC just depends on the model gradients, which can be obtained easily
from any neural network, our active learning algorithm is task-agnostic and mode

l-agnostic. ISAL achieves state-of-the-art performance in different active learning settings for different tasks with different datasets. Compared with previous methods, our method decreases the annotation cost at least by 12%, 13% and 16% on CIFAR10, VOC2012 and COCO, respectively.
****

Visual Alignment Constraint for Continuous Sign Language Recognition
Yuecong Min, Aiming Hao, Xiujuan Chai, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11542-11551
Vision-based Continuous Sign Language Recognition (CSLR) aims to recognize unsegmented signs from image streams. Overfitting is one of the most critical problems in CSLR training, and previous works show that the iterative training scheme can partially solve this problem while also costing more training time. In this study, we revisit the iterative training scheme in recent CSLR works and realize that sufficient training of the feature extractor is critical to solving the overfitting problem. Therefore, we propose a Visual Alignment Constraint (VAC) to enhance the feature extractor with alignment supervision. Specifically, the proposed VAC comprises two auxiliary losses: one focuses on visual features only, and the other enforces prediction alignment between the feature extractor and the alignment module. Moreover, we propose two metrics to reflect overfitting by measuring the prediction inconsistency between the feature extractor and the alignment module. Experimental results on two challenging CSLR datasets show that the proposed VAC makes CSLR networks end-to-end trainable and achieves competitive performance.
****

On the Hidden Treasure of Dialog in Video Question Answering
Deniz Engin, François Schnitzler, Ngoc Q. K. Duong, Yannis Avrithis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2064-2073
High-level understanding of stories in video such as movies and TV shows from raw data is extremely challenging. Modern video question answering (VideoQA) systems often use additional human-made sources like plot synopses, scripts, video descriptions or knowledge bases. In this work, we present a new approach to understand the whole story without such external sources. The secret lies in the dialog: unlike any prior work, we treat dialog as a noisy source to be converted into text description via dialog summarization, much like recent methods treat video. The input of each modality is encoded by transformers independently, and a simple fusion method combines all modalities, using soft temporal attention for localization over long inputs. Our model outperforms the state of the art on the KnowIT VQA dataset by a large margin, without using question-specific human annotation or human-made plot summaries. It even outperforms human evaluators who have never watched any whole episode before. Code is available at https://engindeniz.github.io/dialogsummary-videoqa
****

From Culture to Clothing: Discovering the World Events Behind a Century of Fashion Images
Wei-Lin Hsiao, Kristen Grauman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1066-1075
Fashion is intertwined with external cultural factors, but identifying these links remains a manual process limited to only the most salient phenomena. We propose a data-driven approach to identify specific cultural factors affecting the clothes people wear. Using large-scale datasets of news articles and vintage photos spanning a century, we present a multi-modal statistical model to detect influence relationships between happenings in the world and people's choice of clothing. Furthermore, on two image datasets we apply our model to improve the concrete vision tasks of visual style forecasting and photo timestamping. Our work is a first step towards a computational, scalable, and easily refreshable approach to link culture to clothing.
****

Contextually Plausible and Diverse 3D Human Motion Prediction
Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, Mathieu Salzma

nn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11333-11342
We tackle the task of diverse 3D human motion prediction, that is, forecasting multiple plausible future 3D poses given a sequence of observed 3D poses. In this context, a popular approach consists of using a Conditional Variational Autoencoder (CVAE). However, existing approaches that do so either fail to capture the diversity in human motion, or generate diverse but semantically implausible continuations of the observed motion. In this paper, we address both of these problems by developing a new variational framework that accounts for both diversity and context of the generated future motion. To this end, and in contrast to existing approaches, we condition the sampling of the latent variable that acts as source of diversity on the representation of the past observation, thus encouraging it to carry relevant information. Our experiments demonstrate that our approach yields motions not only of higher quality while retaining diversity, but also that preserve the contextual information contained in the observed motion.
****

BN-NAS: Neural Architecture Search With Batch Normalization
Boyu Chen, Peixia Li, Baopu Li, Chen Lin, Chuming Li, Ming Sun, Junjie Yan, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 307-316
Model training and evaluation are two main time-consuming processes during neural architecture search (NAS). Although weight-sharing based methods have been proposed to reduce the number of trained networks, these methods still need to train the supernet for hundreds of epochs and evaluate thousands of subnets to find the optimal network architecture. In this paper, we propose NAS with Batch Normalization (BN), which we refer to as BN-NAS, to accelerate both the evaluation and training process. For fast evaluation, we propose a novel BN-based indicator that predicts subnet performance at a very early training stage. We further improve the training efficiency by only training the BN parameters during the supernet training. This is based on our observation that training the whole supernet is not necessary while training only BN parameters accelerates network convergence for network architecture search. Extensive experiments show that our method can significantly shorten the time of training supernet by more than 10 times and evaluating subnets by more than 600,000 times without losing accuracy.
****

Condensing a Sequence to One Informative Frame for Video Recognition
Zhaofan Qiu, Ting Yao, Yan Shu, Chong-Wah Ngo, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16311-16320
Video is complex due to large variations in motion and rich content in fine-grained visual details. Abstracting useful information from such information-intensive media requires exhaustive computing resources. This paper studies a two-step alternative that first condenses the video sequence to an informative "frame" and then exploits off-the-shelf image recognition system on the synthetic frame. A valid question is how to define "useful information" and then distill it from a video sequence down to one synthetic frame. This paper presents a novel Informative Frame Synthesis (IFS) architecture that incorporates three objective tasks, i.e., appearance reconstruction, video categorization, motion estimation, and two regularizers, i.e., adversarial learning, color consistency. Each task equips the synthetic frame with one ability, while each regularizer enhances its visual quality. With these, by jointly learning the frame synthesis in an end-to-end manner, the generated frame is expected to encapsulate the required spatio-temporal information useful for video analysis. Extensive experiments are conducted on the large-scale Kinetics dataset. When comparing to baseline methods that map video sequence to a single image, IFS shows superior performance. More remarkably, IFS consistently demonstrates evident improvements on image-based 2D networks and clip-based 3D networks, and achieves comparable performance with the state-of-the-art methods with less computational cost.
****

The Benefit of Distraction: Denoising Camera-Based Physiological Measurements Using Inverse Attention

Ewa M. Nowara, Daniel McDuff, Ashok Veeraraghavan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4955-4964

Attention networks perform well on diverse computer vision tasks. The core idea is that the signal of interest is stronger in some pixels ("foreground"), and by selectively focusing computation on these pixels, networks can extract subtle information buried in noise and other sources of corruption. Our paper is based on one key observation: in many real-world applications, many sources of corruption, such as illumination and motion, are often shared between the "foreground" and the "background" pixels. Can we utilize this to our advantage? We propose the utility of inverse attention networks, which focus on extracting information about these shared sources of corruption. We show that this helps to effectively suppress shared covariates and amplify signal information, resulting in improved performance. We illustrate this on the task of camera-based physiological measurement where the signal of interest is weak and global illumination variations and motion act as significant shared sources of corruption. We perform experiments on three datasets and show that our approach of inverse attention produces state-of-the-art results, increasing the signal-to-noise ratio by up to 5.8 dB, reducing heart rate and breathing rate estimation errors by as much as 30 %, recovering subtle waveform dynamics, and generalizing from RGB to NIR videos without retraining.

****

Collaborative and Adversarial Learning of Focused and Dispersive Representations for Semi-Supervised Polyp Segmentation

Huisi Wu, Guilian Chen, Zhenkun Wen, Jing Qin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3489-3498

Automatic polyp segmentation from colonoscopy images is an essential step in computer aided diagnosis for colorectal cancer. Most of polyp segmentation methods reported in recent years are based on fully supervised deep learning. However, annotation for polyp images by physicians during the diagnosis is time-consuming and costly. In this paper, we present a novel semi-supervised polyp segmentation via collaborative and adversarial learning of focused and dispersive representations learning model, where focused and dispersive extraction module are used to deal with the diversity of location and shape of polyps. In addition, confidence maps produced by a discriminator in an adversarial training framework shows the effectiveness of leveraging unlabeled data and improving the performance of segmentation network. Consistent regularization is further employed to optimize the segmentation networks to strengthen the representation of the outputs of focused and dispersive extraction module. We also propose an auxiliary adversarial learning method to better leverage unlabeled examples to further improve semantic segmentation accuracy. We conduct extensive experiments on two famous polyp datasets: Kvasir-SEG and CVC-Clinic DB. Experimental results demonstrate the effectiveness of the proposed model, consistently outperforming state-of-the-art semi-supervised segmentation models based on adversarial training and even some advanced fully supervised models. Codes will be released upon publication.

****

Active Domain Adaptation via Clustering Uncertainty-Weighted Embeddings

Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, Judy Hoffman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8505-8514

Generalizing deep neural networks to new target domains is critical to their real-world utility. In practice, it may be feasible to get some target data labeled, but to be cost-effective it is desirable to select a maximally-informative subset via active learning (AL). We study the problem of AL under a domain shift, called Active Domain Adaptation (Active DA). We demonstrate how existing AL approaches based solely on model uncertainty or diversity sampling are less effective for Active DA. We propose Clustering Uncertainty-weighted Embeddings (CLUE), a novel label acquisition strategy for Active DA that performs uncertainty-weighted clustering to identify target instances for labeling that are both uncertain under the model and diverse in feature space. CLUE consistently outperforms competing label acquisition strategies for Active DA and AL across learning settings

on 6 diverse domain shifts for image classification.
****

Detail Me More: Improving GAN's Photo-Realism of Complex Scenes
Raghudeep Gadde, Qianli Feng, Aleix M. Martinez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13950-13959
Generative models can synthesize photo-realistic images of a single object. For example, for human faces, algorithms learn to model the local shape and shading of the face components, i.e., changes in the brows, eyes, nose, mouth, jaw line, etc. This is possible because all faces have two brows, two eyes, a nose and a mouth, approximately in the same location. The modeling of complex scenes is however much more challenging because the scene components and their location vary from image to image. For example, living rooms contain a varying number of products belonging to many possible categories and locations, e.g., a lamp may or may not be present in an endless number of possible locations. In the present work, we propose to add a "broker" module in Generative Adversarial Networks (GAN) to solve this problem. The broker is tasked to mediate the use of multiple discriminators in the appropriate image locales. For example, if a lamp is detected or wanted in a specific area of the scene, the broker assigns a fine-grained lamp discriminator to that image patch. This allows the generator to learn the shape and shading models of the lamp. The resulting multi-fine-grained optimization problem is able to synthesize complex scenes with almost the same level of photo-realism as single object images. We demonstrate the generability of the proposed approach on several GAN algorithms (BigGAN, ProGAN, StyleGAN, StyleGAN2), image resolutions (256x256 to 1024x1024), and datasets. Our approach yields significant improvements over state-of-the-art GAN algorithms.
****

Rethinking Self-Supervised Correspondence Learning: A Video Frame-Level Similarity Perspective
Jiarui Xu, Xiaolong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10075-10085
Learning a good representation for space-time correspondence is the key for various computer vision tasks, including tracking object bounding boxes and performing video object pixel segmentation. To learn generalizable representation for correspondence in large-scale, a variety of self-supervised pretext tasks are proposed to explicitly perform object-level or patch-level similarity learning. Instead of following the previous literature, we propose to learn correspondence using Video Frame-level Similarity (VFS) learning, i.e, simply learning from comparing video frames. Our work is inspired by the recent success in image-level contrastive learning and similarity learning for visual recognition. Our hypothesis is that if the representation is good for recognition, it requires the convolutional features to find correspondence between similar objects or parts. Our experiments show surprising results that VFS surpasses state-of-the-art self-supervised approaches for both OTB visual object tracking and DAVIS video object segmentation. We perform detailed analysis on what matters in VFS and reveals new properties on image and frame level similarity learning.
****

Event Stream Super-Resolution via Spatiotemporal Constraint Learning
Siqi Li, Yutong Feng, Yipeng Li, Yu Jiang, Changqing Zou, Yue Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4480-4489
Event cameras are bio-inspired sensors that respond to brightness changes asynchronously and output in the form of event streams instead of frame-based images. They own outstanding advantages compared with traditional cameras: higher temporal resolution, higher dynamic range, and lower power consumption. However, the spatial resolution of existing event cameras is insufficient and challenging to be enhanced at the hardware level while maintaining the asynchronous philosophy of circuit design. Therefore, it is imperative to explore the algorithm of event stream super-resolution, which is a non-trivial task due to the sparsity and strong spatio-temporal correlation of the events from an event camera. In this paper, we propose an end-to-end framework based on spiking neural network for event

stream super-resolution, which can generate high-resolution (HR) event stream fr
om the input low-resolution (LR) event stream. A spatiotemporal constraint learn
ing mechanism is proposed to learn the spatial and temporal distributions of the
 event stream simultaneously. We validate our method on four large-scale dataset
s and the results show that our method achieves state-of-the-art performance. Th
e satisfying results on two downstream applications, i.e. object classification
and image reconstruction, further demonstrate the usability of our method. To pr
ove the application potential of our method, we deploy it on a mobile platform.
The high-quality HR event stream generated by our real-time system demonstrates
the effectiveness and efficiency of our method.
****

PrimitiveNet: Primitive Instance Segmentation With Local Primitive Embedding Und
er Adversarial Metric
Jingwei Huang, Yanfeng Zhang, Mingwei Sun; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 15343-15353
We present PrimitiveNet, a novel approach for high-resolution primitive instance
 segmentation from point clouds on a large scale. Our key idea is to transform t
he global segmentation problem into easier local tasks. We train a high-resoluti
on primitive embedding network to predict explicit geometry features and implici
t latent features for each point. The embedding is jointly trained with an adver
sarial network as a primitive discriminator to decide whether points are from th
e same primitive instance in local neighborhoods. Such local supervision encoura
ges the learned embedding and discriminator to describe local surface properties
 and robustly distinguish different instances. At inference time, network predic
tions are followed by a region growing method to finalize the segmentation. Expe
riments show that our method outperforms existing state-of-the-arts based on mea
n average precision by a significant margin (46.3%) on ABC dataset [??]. We can
process extremely large real scenes covering more than 0.1km^2. Ablation studies
 highlight the contribution of our core designs. Finally, our method can improve
 geometry processing algorithms to abstract scans as lightweight models.
****

FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting
Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang
Wang, Jifeng Dai, Hongsheng Li; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2021, pp. 14040-14049
Transformer, as a strong and flexible architecture for modelling long-range rela
tions, has been widely explored in vision tasks. However, when used in video inp
ainting that requires fine-grained representation, existed method still suffers
from yielding blurry edges in detail due to the hard patch splitting. Here we ai
m to tackle this problem by proposing FuseFormer, a Transformer model designed f
or video inpainting via fine-grained feature fusion based on novel Soft Split an
d Soft Composition operations. The soft split divides feature map into many patc
hes with given overlapping interval. On the contrary, the soft composition opera
tes by stitching different patches into a whole feature map where pixels in over
lapping regions are summed up. These two modules are first used in tokenization
before Transformer layers and de-tokenization after Transformer layers, for effe
ctive mapping between tokens and features. Therefore, sub-patch level informatio
n interaction is enabled for more effective feature propagation between neighbor
ing patches, resulting in synthesizing vivid content for hole regions in videos.
 Moreover, in FuseFormer, we elaborately insert the soft composition and soft sp
lit into the feed-forward network, enabling the 1D linear layers to have the cap
ability of modelling 2D structure. And, the sub-patch level feature fusion abili
ty is further enhanced. In both quantitative and qualitative evaluations, our pr
oposed FuseFormer surpasses state-of-the-art methods. We also conduct detailed a
nalysis to examine its superiority.
****

FASA: Feature Augmentation and Sampling Adaptation for Long-Tailed Instance Segm
entation
Yuhang Zang, Chen Huang, Chen Change Loy; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2021, pp. 3457-3466

Recent methods for long-tailed instance segmentation still struggle on rare obje ct classes with few training data. We propose a simple yet effective method, Fea ture Augmentation and Sampling Adaptation (FASA), that addresses the data scarci ty issue by augmenting the feature space especially for rare classes. Both the F eature Augmentation (FA) and feature sampling components are adaptive to the act ual training status -- FA is informed by the feature mean and variance of observ ed real samples from past iterations, and we sample the generated virtual featur es in a loss-adapted manner to avoid over-fitting. FASA does not require any ela borate loss design, and removes the need for inter-class transfer learning that often involves large cost and manually-defined head/tail class groups. We show F ASA is a fast, generic method that can be easily plugged into standard or long-t ailed segmentation frameworks, with consistent performance gains and little adde d cost. FASA is also applicable to other tasks like long-tailed classification w ith state-of-the-art performance.

****

Online-Trained Upsampler for Deep Low Complexity Video Compression

Jan P. Klopp, Keng-Chi Liu, Shao-Yi Chien, Liang-Gee Chen; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7929-7938

Deep learning for image and video compression has demonstrated promising results both as a standalone technology and a hybrid combination with existing codecs. However, these systems still come with high computational costs. Deep learning m odels are typically applied directly in pixel space, making them expensive when resolutions become large. In this work, we propose an online-trained upsampler t o augment an existing codec. The upsampler is a small neural network trained on an isolated group of frames. Its parameters are signalled to the decoder. This h ybrid solution has a small scope of only 10s or 100s of frames and allows for a low complexity both on the encoding and the decoding side. Our algorithm works i n offline and in zero-latency settings. Our evaluation employs the popular x265 codec on several high-resolution datasets ranging from Full HD to 8K. We demonst rate rate savings between 8.6% and 27.5% and provide ablation studies to show th e impact of our design decisions. In comparison to similar works, our approach p erforms favourably.

****

DTMNet: A Discrete Tchebichef Moments-Based Deep Neural Network for Multi-Focus Image Fusion

Bin Xiao, Haifeng Wu, Xiuli Bi; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 43-51

Compared with traditional methods, the deep learning-based multi-focus image fus ion methods can effectively improve the performance of image fusion tasks. Howev er, the existing deep learning-based methods encounter a common issue of a large number of parameters, which leads to the deep learning models with high time co mplexity and low fusion efficiency. To address this issue, we propose a novel di screte Tchebichef moment-based Deep neural network, termed as DTMNet, for multi- focus image fusion. The proposed DTMNet is an end-to-end deep neural network wit h only one convolutional layer and three fully connected layers. The convolution al layer is fixed with DTM coefficients (DTMConv) to extract high/low-frequency information without learning parameters effectively. The three fully connected l ayers have learnable parameters for feature classification. Therefore, the propo sed DTMNet for multi-focus image fusion has a small number of parameters (0.01M paras vs. 4.93M paras of regular CNN) and high computational efficiency (0.32s v s. 79.09s by regular CNN to fuse an image). In addition, a large-scale multi-foc us image dataset is synthesized for training and verifying the deep learning mod el. Experimental results on three public datasets demonstrate that the proposed method is competitive with or even outperforms the state-of-the-art multi-focus image fusion methods in terms of subjective visual perception and objective eval uation metrics.

****

Interactive Prototype Learning for Egocentric Action Recognition

Xiaohan Wang, Linchao Zhu, Heng Wang, Yi Yang; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 8168-8177

Egocentric video recognition is a challenging task that requires to identify both the actor's motion and the active object that the actor interacts with. Recognizing the active object is particularly hard due to the cluttered background with distracting objects, the frequent field of view changes, severe occlusion, etc. To improve the active object classification, most existing methods use object detectors or human gaze information, which are computationally expensive or require labor-intensive annotations. To avoid these additional costs, we propose an end-to-end Interactive Prototype Learning (IPL) framework to learn better active object representations by leveraging the motion cues from the actor. First, we introduce a set of verb prototypes to disentangle active object features from distracting object features. Each prototype corresponds to a primary motion pattern of an egocentric action, offering a distinctive supervision signal for active object feature learning. Second, we design two interactive operations to enable the extraction of active object features, i.e., noun-to-verb assignment and verb-to-noun selection. These operations are parameter-efficient and can learn judicious location-aware features on top of 3D CNN backbones. We demonstrate that the IPL framework can generalize to different backbones and outperform the state-of-the-art on three large-scale egocentric video datasets, i.e., EPIC-KITCHENS-55, EPIC-KITCHENS-100 and EGTEA.

****

MBA-VO: Motion Blur Aware Visual Odometry

Peidong Liu, Xingxing Zuo, Viktor Larsson, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5550-5559

Motion blur is one of the major challenges remaining for visual odometry methods. In low-light conditions where longer exposure times are necessary, motion blur can appear even for relatively slow camera motions. In this paper we present a novel hybrid visual odometry pipeline with direct approach that explicitly models and estimates the camera's local trajectory within exposure time. This allows us to actively compensate for any motion blur that occurs due to the camera motion. In addition, we also contribute a novel benchmarking dataset for motion blur aware visual odometry. In experiments we show that by directly modeling the image formation process we are able to improve robustness of the visual odometry, while keeping comparable accuracy as that for images without motion blur.

****

Co2L: Contrastive Continual Learning

Hyuntak Cha, Jaeho Lee, Jinwoo Shin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9516-9525

Recent breakthroughs in self-supervised learning show that such algorithms learn visual representations that can be transferred better to unseen tasks than cross-entropy based methods which rely on task-specific supervision. In this paper, we found that the similar holds in the continual learning context: contrastively learned representations are more robust against the catastrophic forgetting than ones trained with the cross-entropy objective. Based on this novel observation, we propose a rehearsal-based continual learning algorithm that focuses on continually learning and maintaining transferable representations. More specifically, the proposed scheme (1) learns representations using the contrastive learning objective, and (2) preserves learned representations using a self-supervised distillation step. We conduct extensive experimental validations under popular benchmark image classification datasets, where our method sets the new state-of-the-art performance. Source code is available at https://github.com/chaht01/Co2L.

****

STR-GQN: Scene Representation and Rendering for Unknown Cameras Based on Spatial Transformation Routing

Wen-Cheng Chen, Min-Chun Hu, Chu-Song Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5966-5975

Geometry-aware modules are widely applied in recent deep learning architectures for scene representation and rendering. However, these modules require intrinsic camera information that might not be obtained accurately. In this paper, we propose a Spatial Transformation Routing (STR) mechanism to model the spatial properties without applying any geometric prior. The STR mechanism treats the spatial

transformation as the message passing process, and the relation between the view poses and the routing weights is modeled by an end-to-end trainable neural network. Besides, an Occupancy Concept Mapping (OCM) framework is proposed to provide explainable rationals for scene-fusion processes. We conducted experiments on several datasets and show that the proposed STR mechanism improves the performance of the Generative Query Network (GQN). The visualization results reveal that the routing process can pass the observed information from one location of some view to the associated location in the other view, which demonstrates the advantage of the proposed model in terms of spatial cognition.
****

Unconstrained Scene Generation With Locally Conditioned Radiance Fields
Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, Joshua M. Susskind; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14304-14313
We tackle the challenge of learning a distribution over complex, realistic, indoor scenes. In this paper, we introduce Generative Scene Networks (GSN), which learns to decompose scenes into a collection of many local radiance fields that can be rendered from a free moving camera. Our model can be used as a prior to generate new scenes, or to complete a scene given only sparse 2D observations. Recent work has shown that generative models of radiance fields can capture properties such as multi-view consistency and view-dependent lighting. However, these models are specialized for constrained viewing of single objects, such as cars or faces. Due to the size and complexity of realistic indoor environments, existing models lack the representational capacity to adequately capture them. Our decomposition scheme scales to larger and more complex scenes while preserving details and diversity, and the learned prior enables high-quality rendering from view-points that are significantly different from observed viewpoints. When compared to existing models, GSN produces quantitatively higher quality scene renderings across several different scene datasets.
****

3D Human Pose Estimation With Spatial and Temporal Transformers
Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, Zhengming Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11656-11665
Transformer architectures have become the model of choice in natural language processing and are now being introduced into computer vision tasks such as image classification, object detection, and semantic segmentation. However, in the field of human pose estimation, convolutional architectures still remain dominant. In this work, we present PoseFormer, a purely transformer-based approach for 3D human pose estimation in videos without convolutional architectures involved. Inspired by recent developments in vision transformers, we design a spatial-temporal transformer structure to comprehensively model the human joint relations within each frame as well as the temporal correlations across frames, then output an accurate 3D human pose of the center frame. We quantitatively and qualitatively evaluate our method on two popular and standard benchmark datasets: Human3.6M and MPI-INF-3DHP. Extensive experiments show that PoseFormer achieves state-of-the-art performance on both datasets. Our code and model will be publicly available.
****

Self-Supervised Representation Learning From Flow Equivariance
Yuwen Xiong, Mengye Ren, Wenyuan Zeng, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10191-10200
Self-supervised representation learning is able to learn semantically meaningful features; however, much of its recent success relies on multiple crops of an image with very few objects. Instead of learning view-invariant representation from simple images, humans learn representations in a complex world with changing scenes by observing object movement, deformation, pose variation, and ego motion. Motivated by this ability, we present a new self-supervised learning representation framework that can be directly deployed on a video stream of complex scenes with many moving objects. Our framework features a simple flow equivariance obj

ective that encourages the network to predict the features of another frame by a
pplying a flow transformation to the features of the current frame. Our represen
tations, learned from high-resolution raw video, can be readily used for downstr
eam tasks on static images. Readout experiments on challenging semantic segmenta
tion, instance segmentation, and object detection benchmarks show that we are ab
le to outperform representations obtained from previous state-of-the-art methods
 including SimCLR and BYOL.
****

Continual Learning for Image-Based Camera Localization
Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Juho Kannala; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
 pp. 3252-3262
For several emerging technologies such as augmented reality, autonomous driving
and robotics, visual localization is a critical component. Directly regressing c
amera pose/3D scene coordinates from the input image using deep neural networks
has shown great potential. However, such methods assume a stationary data distri
bution with all scenes simultaneously available during training. In this paper,
we approach the problem of visual localization in a continual learning setup --
whereby the model is trained on scenes in an incremental manner. Our results sho
w that similar to the classification domain, non-stationary data induces catastr
ophic forgetting in deep networks for visual localization. To address this issue
, a strong baseline based on storing and replaying images from a fixed buffer is
 proposed. Furthermore, we propose a new sampling method based on coverage score
 (Buff-CS) that adapts the existing sampling strategies in the buffering process
 to the problem of visual localization. Results demonstrate consistent improveme
nts over standard buffering methods on two challenging datasets -- 7Scenes, 12Sc
enes, and also 19Scenes by combining the former scenes.
****

Visual-Textual Attentive Semantic Consistency for Medical Report Generation
Yi Zhou, Lei Huang, Tao Zhou, Huazhu Fu, Ling Shao; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2021, pp. 3985-3994
Diagnosing diseases from medical radiographs and writing reports requires profes
sional knowledge and is time-consuming. To address this, automatic medical repor
t generation approaches have recently gained interest. However, identifying dise
ases as well as correctly predicting their corresponding sizes, locations and ot
her medical description patterns, which is essential for generating high-quality
 reports, is challenging. Although previous methods focused on producing readabl
e reports, how to accurately detect and describe findings that match with the qu
ery X-Ray has not been successfully addressed. In this paper, we propose a multi
-modality semantic attention model to integrate visual features, predicted key f
inding embeddings, as well as clinical features, and progressively decode report
s with visual-textual semantic consistency. First, multi-modality features are e
xtracted and attended with the hidden states from the sentence decoder, to encod
e enriched context vectors for better decoding a report. These modalities includ
e regional visual features of scans, semantic word embeddings of the top-K findi
ngs predicted with high probabilities, and clinical features of indications. Sec
ond, the progressive report decoder consists of a sentence decoder and a word de
coder, where we propose image-sentence matching and description accuracy losses
to constrain the visual-textual semantic consistency. Extensive experiments on t
he public MIMIC-CXR and IU X-Ray datasets show that our model achieves consisten
t improvements over the state-of-the-art methods.
****

Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective With
Transformers
Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russ
ell H. Taylor, Mathias Unberath; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2021, pp. 6197-6206
Stereo depth estimation relies on optimal correspondence matching between pixels
 on epipolar lines in the left and right images to infer depth. In this work, we
 revisit the problem from a sequence-to-sequence correspondence perspective to r

eplace cost volume construction with dense pixel matching using position informa
tion and attention. This approach, named STereo TRansformer (STTR), has several
advantages: It 1) relaxes the limitation of a fixed disparity range, 2) identifi
es occluded regions and provides confidence estimates, and 3) imposes uniqueness
 constraints during the matching process. We report promising results on both sy
nthetic and real-world datasets and demonstrate that STTR generalizes across dif
ferent domains, even without fine-tuning.
****

Augmenting Depth Estimation With Geospatial Context
Scott Workman, Hunter Blanton; Proceedings of the IEEE/CVF International Confere
nce on Computer Vision (ICCV), 2021, pp. 4562-4571
Modern cameras are equipped with a wide array of sensors that enable recording t
he geospatial context of an image. Taking advantage of this, we explore depth es
timation under the assumption that the camera is geocalibrated, a problem we ref
er to as geo-enabled depth estimation. Our key insight is that if capture locati
on is known, the corresponding overhead viewpoint offers a valuable resource for
 understanding the scale of the scene. We propose an end-to-end architecture for
 depth estimation that uses geospatial context to infer a synthetic ground-level
 depth map from a co-located overhead image, then fuses it inside of an encoder/
decoder style segmentation network. To support evaluation of our methods, we ext
end a recently released dataset with overhead imagery and corresponding height m
aps. Results demonstrate that integrating geospatial context significantly reduc
es error compared to baselines, both at close ranges and when evaluating at much
 larger distances than existing benchmarks consider.
****

Explaining Local, Global, and Higher-Order Interactions in Deep Learning
Samuel Lerman, Charles Venuto, Henry Kautz, Chenliang Xu; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1224-1233
We present a simple yet highly generalizable method for explaining interacting p
arts within a neural network's reasoning process. First, we design an algorithm
based on cross derivatives for computing statistical interaction effects between
 individual features, which is generalized to both 2-way and higher-order (3-way
 or more) interactions. We present results side by side with a weight-based attr
ibution technique, corroborating that cross derivatives are a superior metric fo
r both 2-way and higher-order interaction detection. Moreover, we extend the use
 of cross derivatives as an explanatory device in neural networks to the compute
r vision setting by expanding Grad-CAM, a popular gradient-based explanatory too
l for CNNs, to the higher order. While Grad-CAM can only explain the importance
of individual objects in images, our method, which we call Taylor-CAM, can expla
in a neural network's relational reasoning across multiple objects. We show the
success of our explanations both qualitatively and quantitatively, including wit
h a user study. We will release all code as a tool package to facilitate explain
able deep learning.
****

Learning Attribute-Driven Disentangled Representations for Interactive Fashion R
etrieval
Yuxin Hou, Eleonora Vig, Michael Donoser, Loris Bazzani; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12147-12157
Interactive retrieval for online fashion shopping provides the ability of changi
ng image retrieval results according to the user feedback. One common problem in
 interactive retrieval is that a specific user interaction (e.g., changing the c
olor of a T-shirt) causes other aspects to change inadvertently (e.g., the resul
ts have a sleeve type different from that of the query). This is a consequence o
f existing methods learning visual representations that are entangled in the emb
edding space, which limits the controllability of the retrieved results. We prop
ose to leverage on the semantics of visual attributes to train convolutional net
works that learn attribute-specific subspaces for each attribute type to obtain
disentangled representations. Operations, such as swapping out a particular attr
ibute value for another, impact the attribute at hand and leave others untouched
. We show that our model can be tailored to deal with different retrieval tasks

while maintaining its disentanglement property. We obtained state-of-the-art performance on three interactive fashion retrieval tasks: attribute manipulation retrieval, conditional similarity retrieval, and outfit complementary item retrieval. We will make code and models publicly available.
****

SemiHand: Semi-Supervised Hand Pose Estimation With Consistency
Linlin Yang, Shicheng Chen, Angela Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11364-11373
We present SemiHand, a semi-supervised framework for 3D hand pose estimation from monocular images. We pre-train the model on labelled synthetic data and fine-tune it on unlabelled real-world data by pseudo-labeling with consistency training. By design, we introduce data augmentation of differing difficulties, consistency regularizer, label correction and sample selection for RGB-based 3D hand pose estimation. In particular, by approximating the hand masks from hand poses, we propose a cross-modal consistency and leverage semantic predictions to guide the predicted poses. Meanwhile, we introduce pose registration as label correction to guarantee the biomechanical feasibility of hand bone lengths. Experiments show that our method achieves a favorable improvement on real-world datasets after fine-tuning.
****

Efficient Action Recognition via Dynamic Knowledge Propagation
Hanul Kim, Mihir Jain, Jun-Tae Lee, Sungrack Yun, Fatih Porikli; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13719-13728
Efficient action recognition has become crucial to extend the success of action recognition to many real-world applications. Contrary to most existing methods, which mainly focus on selecting salient frames to reduce the computation cost, we focus more on making the most of the selected frames. To this end, we employ two networks of different capabilities that operate in tandem to efficiently recognize actions. Given a video, the lighter network processes more frames while the heavier one only processes a few. In order to enable the effective interaction between the two, we propose dynamic knowledge propagation based on a cross-attention mechanism. This is the main component of our framework that is essentially a student-teacher architecture, but as the teacher model continues to interact with the student model during inference, we call it a dynamic student-teacher framework. Through extensive experiments, we demonstrate the effectiveness of each component of our framework. Our method outperforms competing state-of-the-art methods on two video datasets: ActivityNet-v1.3 and Mini-Kinetics.
****

Bias Loss for Mobile Neural Networks
Lusine Abrahamyan, Valentin Ziatchin, Yiming Chen, Nikos Deligiannis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6556-6566
Compact convolutional neural networks (CNNs) have witnessed exceptional improvements in performance in recent years. However, they still fail to provide the same predictive power as CNNs with a large number of parameters. The diverse and even abundant features captured by the layers is an important characteristic of these successful CNNs. However, differences in this characteristic between large CNNs and their compact counterparts have rarely been investigated. In compact CNNs, due to the limited number of parameters, abundant features are unlikely to be obtained, and feature diversity becomes an essential characteristic. Diverse features present in the activation maps derived from a data point during model inference may indicate the presence of a set of unique descriptors necessary to distinguish between objects of different classes. In contrast, data points with low feature diversity may not provide a sufficient amount of unique descriptors to make a valid prediction; we refer to them as random predictions. Random predictions can negatively impact the optimization process and harm the final performance. This paper proposes addressing the problem raised by random predictions by reshaping the standard cross-entropy to make it biased toward data points with a limited number of unique descriptive features. Our novel Bias Loss focuses the tr

aining on a set of valuable data points and prevents the vast number of samples with poor learning features from misleading the optimization process. Furthermore, to show the importance of diversity, we present a family of SkipblockNet models whose architectures are brought to boost the number of unique descriptors in the last layers. Experiments conducted on benchmark datasets demonstrate the superiority of the proposed loss function over the cross-entropy loss. Moreover, our SkipblockNet-M can achieve 1% higher classification accuracy than MobileNetV3 Large with similar computational cost on the ImageNet ILSVRC-2012 classification dataset. The code is available on the link - https://github.com/lusinlu/biasloss_skipblocknet.

****

Visual Scene Graphs for Audio Source Separation

Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, Anoop Cherian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1204-1213

State-of-the-art approaches for visually-guided audio source separation typically assume sources that have characteristic sounds, such as musical instruments. These approaches often ignore the visual context of these sound sources or avoid modeling object interactions that may be useful to better characterize the sources, especially when the same object class may produce varied sounds from distinct interactions. To address this challenging problem, we propose Audio Visual Scene Graph Segmenter (AVSGS), a novel deep learning model that embeds the visual structure of the scene as a graph and segments this graph into subgraphs, each subgraph being associated with a unique sound obtained by co-segmenting the audio spectrogram. At its core, AVSGS uses a recursive neural network that emits mutually-orthogonal sub-graph embeddings of the visual graph using multi-head attention. These embeddings are used for conditioning an audio encoder-decoder towards source separation. Our pipeline is trained end-to-end via a self-supervised task consisting of separating audio sources using the visual graph from artificially mixed sounds. In this paper, we also introduce an ""in the wild" video dataset for sound source separation that contains multiple non-musical sources, which we call Audio Separation in the Wild (ASIW). This dataset is adapted from the AudioCaps dataset, and provides a challenging, natural, and daily-life setting for source separation. Thorough experiments on the proposed ASIW and the standard MUSIC datasets demonstrate state-of-the-art sound separation performance of our method against recent prior approaches.

****

Beyond Trivial Counterfactual Explanations With Diverse Valuable Explanations

Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, David Vazquez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1056-1065

Explainability for machine learning models has gained considerable attention within the research community given the importance of deploying more reliable machine-learning systems. In computer vision applications, generative counterfactual methods indicate how to perturb a model's input to change its prediction, providing details about the model's decision-making. Current methods tend to generate trivial counterfactuals about a model's decisions, as they often suggest to exaggerate or remove the presence of the attribute being classified. For the machine learning practitioner, these types of counterfactuals offer little value, since they provide no new information about undesired model or data biases. In this work, we identify the problem of trivial counterfactual generation and we propose DiVE to alleviate it. DiVE learns a perturbation in a disentangled latent space that is constrained using a diversity-enforcing loss to uncover multiple valuable explanations about the model's prediction. Further, we introduce a mechanism to prevent the model from producing trivial explanations. Experiments on CelebA and Synbols demonstrate that our model improves the success rate of producing high-quality valuable explanations when compared to previous state-of-the-art methods.

****

Homogeneous Architecture Augmentation for Neural Predictor

Yuqiao Liu, Yehui Tang, Yanan Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12249-12258
Neural Architecture Search (NAS) can automatically design well-performed architectures of Deep Neural Networks (DNNs) for the tasks at hand. However, one bottleneck of NAS is the prohibitively computational cost largely due to the expensive performance evaluation. The neural predictors can directly estimate the performance without any training of the DNNs to be evaluated, thus have drawn increasing attention from researchers. Despite their popularity, they also suffer a severe limitation: the shortage of annotated DNN architectures for effectively training the neural predictors. In this paper, we proposed Homogeneous Architecture Augmentation for Neural Predictor (HAAP) of DNN architectures to address the issue aforementioned. Specifically, a homogeneous architecture augmentation algorithm is proposed in HAAP to generate sufficient training data taking the use of homogeneous representation. Furthermore, the one-hot encoding strategy is introduced into HAAP to make the representation of DNN architectures more effective. The experiments have been conducted on both NAS-Benchmark-101 and NAS-Bench-201 dataset. The experimental results demonstrate that the proposed HAAP algorithm outperforms the state of the arts compared, yet with much less training data. In addition, the ablation studies on both benchmark datasets have also shown the universality of the homogeneous architecture augmentation. Our code has been made available at https://github.com/lyq998/HAAP.
****

Co-Scale Conv-Attentional Image Transformers
Weijian Xu, Yifan Xu, Tyler Chang, Zhuowen Tu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9981-9990
In this paper, we present Co-scale conv-attentional image Transformers (CoaT), a Transformer-based image classifier equipped with co-scale and conv-attentional mechanisms. First, the co-scale mechanism maintains the integrity of Transformers' encoder branches at individual scales, while allowing representations learned at different scales to effectively communicate with each other; we design a series of serial and parallel blocks to realize the co-scale mechanism. Second, we devise a conv-attentional mechanism by realizing a relative position embedding formulation in the factorized attention module with an efficient convolution-like implementation. CoaT empowers image Transformers with enriched multi-scale and contextual modeling capabilities. On ImageNet, relatively small CoaT models attain superior classification results compared with similar-sized convolutional neural networks and image/vision Transformers. The effectiveness of CoaT's backbone is also illustrated on object detection and instance segmentation, demonstrating its applicability to downstream computer vision tasks.
****

Impact of Aliasing on Generalization in Deep Convolutional Networks
Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, Ross Goroshin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10529-10538
We investigate the impact of aliasing on generalization in Deep Convolutional Networks and show that data augmentation schemes alone are unable to prevent it due to structural limitations in widely used architectures. Drawing insights from frequency analysis theory, we take a closer look at Resnet and EfficientNet architectures and review the trade-off between aliasing and information loss in each of their major components. We show how to mitigate aliasing by inserting non-trainable low-pass filters at key locations, particularly where networks lack the capacity to learn them. These simple architectural changes lead to substantial improvements in generalization on i.i.d. and even more on out-of-distribution conditions, such as image classification under natural corruptions on ImageNet-C and few-shot learning on Meta-Dataset. State-of-the art results are achieved on both datasets without introducing additional trainable parameters and using the default hyper-parameters of open source codebases.
****

PARE: Part Attention Regressor for 3D Human Body Estimation
Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, Michael J. Black; Proceedin

gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11127-11137
Despite significant progress, we show that state of the art 3D human pose and shape estimation methods remain sensitive to partial occlusion and can produce dramatically wrong predictions although much of the body is observable. To address this, we introduce a soft attention mechanism, called the Part Attention REgressor (PARE), that learns to predict body-part-guided attention masks. We observe that state-of-the-art methods rely on global feature representations, making them sensitive to even small occlusions. In contrast, PARE's part-guided attention mechanism overcomes these issues by exploiting information about the visibility of individual body parts while leveraging information from neighboring body-parts to predict occluded parts. We show qualitatively that PARE learns sensible attention masks, and quantitative evaluation confirms that PARE achieves more accurate and robust reconstruction results than existing approaches on both occlusion-specific and standard benchmarks. The code and data are available for research purposes at https://pare.is.tue.mpg.de/
****

RePOSE: Fast 6D Object Pose Refinement via Deep Texture Rendering
Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, Kris M. Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3303-3312
We present RePOSE, a fast iterative refinement method for 6D object pose estimation. Prior methods perform refinement by feeding zoomed-in input and rendered RGB images into a CNN and directly regressing an update of a refined pose. Their runtime is slow due to the computational cost of CNN, which is especially prominent in multiple-object pose refinement. To overcome this problem, RePOSE leverages image rendering for fast feature extraction using a 3D model with a learnable texture. We call this deep texture rendering, which uses a shallow multi-layer perceptron to directly regress a view-invariant image representation of an object. Furthermore, we utilize differentiable Levenberg-Marquard (LM) optimization to refine a pose fast and accurately by minimizing the feature-metric error between the input and rendered image representations without the need of zooming in. These image representations are trained such that differentiable LM optimization converges within few iterations. Consequently, RePOSE runs at 92 FPS and achieves state-of-the-art accuracy of 51.6% on the Occlusion LineMOD dataset - a 4.1% absolute improvement over the prior art, and comparable result on the YCB-Video dataset with a much faster runtime. The code is available at https://github.com/sh8/repose.
****

How To Design a Three-Stage Architecture for Audio-Visual Active Speaker Detection in the Wild
Okan Köpüklü, Maja Taseska, Gerhard Rigoll; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1193-1203
Successful active speaker detection requires a three-stage pipeline: (i) audio-visual encoding for all speakers in the clip, (ii) inter-speaker relation modeling between a reference speaker and the background speakers within each frame, and (iii) temporal modeling for the reference speaker. Each stage of this pipeline plays an important role for the final performance of the created architecture. Based on a series of controlled experiments, this work presents several practical guidelines for audio-visual active speaker detection. Correspondingly, we present a new architecture called ASDNet, which achieves a new state-of-the-art on the AVA-ActiveSpeaker dataset with a mAP of 93.5% outperforming the second best with a large margin of 4.7%. Our code and pretrained models are publicly available.
****

Do Different Deep Metric Learning Losses Lead to Similar Learned Features?
Konstantin Kobs, Michael Steininger, Andrzej Dulny, Andreas Hotho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10644-10654
Recent studies have shown that many deep metric learning loss functions perform

very similarly under the same experimental conditions. One potential reason for this unexpected result is that all losses let the network focus on similar image regions or properties. In this paper, we investigate this by conducting a two-step analysis to extract and compare the learned visual features of the same model architecture trained with different loss functions: First, we compare the learned features on the pixel level by correlating saliency maps of the same input images. Second, we compare the clustering of embeddings for several image properties, e.g. object color or illumination. To provide independent control over these properties, photo-realistic 3D car renders similar to images in the Cars196 dataset are generated. In our analysis, we compare 14 pretrained models from a recent study and find that, even though all models perform similarly, different loss functions can guide the model to learn different features. We especially find differences between classification and ranking based losses. Our analysis also shows that some seemingly irrelevant properties can have significant influence on the resulting embedding. We encourage researchers from the deep metric learning community to use our methods to get insights into the features learned by their proposed methods.

****

Just Ask: Learning To Answer Questions From Millions of Narrated Videos
Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1686-1697

Recent methods for visual question answering rely on large-scale annotated datasets. Manual annotation of questions and answers for videos, however, is tedious, expensive and prevents scalability. In this work, we propose to avoid manual annotation and generate a large-scale training dataset for video question answering making use of automatic cross-modal supervision. We leverage a question generation transformer trained on text data and use it to generate question-answer pairs from transcribed video narrations. Given narrated videos, we then automatically generate the HowToVQA69M dataset with 69M video-question-answer triplets. To handle the open vocabulary of diverse answers in this dataset, we propose a training procedure based on a contrastive loss between a video-question multi-modal transformer and an answer transformer. We introduce the zero-shot VideoQA task and show excellent results, in particular for rare answers. Furthermore, we demonstrate our method to significantly outperform the state of the art on MSRVTT-QA, MSVD-QA, ActivityNet-QA and How2QA. Finally, for a detailed evaluation we introduce iVQA, a new VideoQA dataset with reduced language biases and high-quality redundant manual annotations.

****

Towards Face Encryption by Generating Adversarial Identity Masks
Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, Hui Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3897-3907

As billions of personal data being shared through social media and network, the data privacy and security have drawn an increasing attention. Several attempts have been made to alleviate the leakage of identity information from face photos, with the aid of, e.g., image obfuscation techniques. However, most of the present results are either perceptually unsatisfactory or ineffective against face recognition systems. Our goal in this paper is to develop a technique that can encrypt the personal photos such that they can protect users from unauthorized face recognition systems but remain visually identical to the original version for human beings. To achieve this, we propose a targeted identity-protection iterative method (TIP-IM) to generate adversarial identity masks which can be overlaid on facial images, such that the original identities can be concealed without sacrificing the visual quality. Extensive experiments demonstrate that TIP-IM provides 95%+ protection success rate against various state-of-the-art face recognition models under practical open-set test scenarios. Besides, we also show the practical and effective applicability of our method on a commercial API service.

****

UniT: Multimodal Multitask Learning With a Unified Transformer

Ronghang Hu, Amanpreet Singh; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2021, pp. 1439-1449
We propose UniT, a Unified Transformer model to simultaneously learn the most pr
ominent tasks across different domains, ranging from object detection to natural
 language understanding and multimodal reasoning. Based on the transformer encod
er-decoder architecture, our UniT model encodes each input modality with an enco
der and makes predictions on each task with a shared decoder over the encoded in
put representations, followed by task-specific output heads. The entire model is
 jointly trained end-to-end with losses from each task. Compared to previous eff
orts on multi-task learning with transformers, we share the same model parameter
s across all tasks instead of separately fine-tuning task-specific models and ha
ndle a much higher variety of tasks across different domains. In our experiments
, we learn 7 tasks jointly over 8 datasets, achieving strong performance on each
 task with significantly fewer parameters. Our code is available in MMF at https
://mmf.sh.
****

CSG-Stump: A Learning Friendly CSG-Like Representation for Interpretable Shape P
arsing
Daxuan Ren, Jianmin Zheng, Jianfei Cai, Jiatong Li, Haiyong Jiang, Zhongang Cai,
 Junzhe Zhang, Liang Pan, Mingyuan Zhang, Haiyu Zhao, Shuai Yi; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12478-
12487
Generating an interpretable and compact representation of 3D shapes from point c
louds is an important and challenging problem. This paper presents CSG-Stump Net
, an unsupervised end-to-end network for learning shapes from point clouds and d
iscovering the underlying constituent modeling primitives and operations as well
. At the core is a three-level structure called CSG-Stump , consisting of a co
mplement layer at the bottom, an intersection layer in the middle, and a union l
ayer at the top. CSG-Stump is proven to be equivalent to CSG in terms of represe
ntation, therefore inheriting the interpretable, compact and editable nature of
CSG while freeing from CSG's complex tree structures. Particularly, the CSG-Stum
p has a simple and regular structure, allowing neural networks to give outputs o
f a constant dimensionality, which makes itself deep-learning friendly. Due to t
hese characteristics of CSG-Stump, CSG-Stump Net achieves superior results compa
red to previous CSG-based methods and generates much more appealing shapes, as c
onfirmed by extensive experiment
****

Compressing Visual-Linguistic Model via Knowledge Distillation
Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, Zicheng Liu;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2021, pp. 1428-1438
Despite exciting progress in pre-training for visual-linguistic (VL) representat
ions, very few aspire to a small VL model. In this paper, we study knowledge dis
tillation(KD) to effectively compress a transformer-based large VL model into a
small VL model. The major challenge arises from the inconsistent regional visual
 tokens extracted from different detectors of Teacher and Student, resulting in
the misalignment of hidden representations and attention distributions. To addre
ss the problem, we retrain and adapt the Teacher by using the same region propos
als from Student's detector while the features are from Teacher's own object det
ector. With aligned network inputs, the adapted Teacher is capable of transferri
ng the knowledge through the intermediate representations. Specifically, we use
the mean square error loss to mimic the attention distribution inside the transf
ormer block and present a token-wise noise contrastive loss to align the hidden
state by contrasting with negative representations stored in a sample queue. To
this end, we show that our proposed distillation significantly improves the perf
ormance of small VL models on image captioning and visual question answering tas
ks. It reaches 120.8 in CIDEr score on COCO captioning, an improvement of 5.1 ov
er its non-distilled counterpart; and an accuracy of 69.8 on VQA 2.0, a 0.8 gain
 from the baseline. Our extensive experiments and ablations confirm the effectiv
e-ness of VL distillation in both pre-training and fine-tuning stages.

****
Full-Duplex Strategy for Video Object Segmentation
Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4922-4933
Appearance and motion are two important sources of information in video object segmentation (VOS). Previous methods mainly focus on using simplex solutions, lowering the upper bound of feature collaboration among and across these two cues. In this paper, we study a novel framework, termed the FSNet (Full-duplex Strategy Network), which designs a relational cross-attention module (RCAM) to achieve the bidirectional message propagation across embedding subspaces. Furthermore, the bidirectional purification module (BPM) is introduced to update the inconsistent features between the spatial-temporal embeddings, effectively improving the model robustness. By considering the mutual restraint within the full-duplex strategy, our FSNet performs the cross-modal feature-passing (i.e., transmission and receiving) simultaneously before the fusion and decoding stage, making it robust to various challenging scenarios (e.g., motion blur, occlusion) in VOS. Extensive experiments on five popular benchmarks (i.e., DAVIS16, FBMS, MCL, SegTrack-V2, and DAVSOD19) show that our FSNet outperforms other state-of-the-arts for both the VOS and video salient object detection tasks.
****
Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments With Support Samples
Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, Michael Rabbat; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8443-8452
This paper proposes a novel method of learning by predicting view assignments with support samples (PAWS). The method trains a model to minimize a consistency loss, which ensures that different views of the same unlabeled instance are assigned similar pseudo-labels. The pseudo-labels are generated non-parametrically, by comparing the representations of the image views to those of a set of randomly sampled labeled images. The distance between the view representations and labeled representations is used to provide a weighting over class labels, which we interpret as a soft pseudo-label. By non-parametrically incorporating labeled samples in this way, PAWS extends the distance-metric loss used in self-supervised methods such as BYOL and SwAV to the semi-supervised setting. Despite the simplicity of the approach, PAWS outperforms other semi-supervised methods across architectures, setting a new state-of-the-art for a ResNet-50 on ImageNet trained with either 10% or 1% of the labels, reaching 75% and 66% top-1 respectively. This is achieved with only 200 epochs of training, which is 4x less than the previous best method.
****
Unsupervised Non-Rigid Image Distortion Removal via Grid Deformation
Nianyi Li, Simron Thapa, Cameron Whyte, Albert W. Reed, Suren Jayasuriya, Jinwei Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2522-2532
Many computer vision problems face difficulties when imaging through turbulent refractive media (e.g., air and water) due to the refraction and scattering of light. These effects cause geometric distortion that requires either handcrafted physical priors or supervised learning methods to remove. In this paper, we present a novel unsupervised network to recover the latent distortion-free image. The key idea is to model non-rigid distortions as deformable grids. Our network consists of a grid deformer that estimates the distortion field and an image generator that outputs the distortion-free image. By leveraging the positional encoding operator, we can simplify the network structure while maintaining fine spatial details in the recovered images. Our method doesn't need to be trained on labeled data and has good transferability across various turbulent image datasets with different types of distortions. Extensive experiments on both simulated and real-captured turbulent images demonstrate that our method can remove both air and water distortions without much customization.

****
Stochastic Transformer Networks With Linear Competing Units: Application To End-to-End SL Translation
Andreas Voskou, Konstantinos P. Panousis, Dimitrios Kosmopoulos, Dimitris N. Metaxas, Sotirios Chatzis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11946-11955
Automating sign language translation (SLT) is a challenging real-world application. Despite its societal importance, though, research progress in the field remains rather poor. Crucially, existing methods that yield viable performance necessitate the availability of laborious to obtain gloss sequence groundtruth. In this paper, we attenuate this need, by introducing an end-to-end SLT model that does not entail explicit use of glosses; the model only needs text groundtruth. This is in stark contrast to existing end-to-end models that use gloss sequence groundtruth, either in the form of a modality that is recognized at an intermediate model stage, or in the form of a parallel output process, jointly trained with the SLT model. Our approach constitutes a Transformer network with a novel type of layers that combines: (i) local winner-takes-all (LWTA) layers with stochastic winner sampling, instead of conventional ReLU layers, (ii) stochastic weights with posterior distributions estimated via variational inference, and (iii) a weight compression technique at inference time that exploits estimated posterior variance to perform massive, almost lossless compression. We demonstrate that our approach can reach the currently best reported BLEU-4 score on the PHOENIX 2014T benchmark, but without making use of glosses for model training, and with a memory footprint reduced by more than 70%.
****
BlockCopy: High-Resolution Video Processing With Block-Sparse Feature Propagation and Online Policies
Thomas Verelst, Tinne Tuytelaars; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5158-5167
In this paper we propose BlockCopy, a scheme that accelerates pretrained frame-based CNNs to process video more efficiently, compared to standard frame-by-frame processing. To this end, a lightweight policy network determines important regions in an image, and operations are applied on selected regions only, using custom block-sparse convolutions. Features of non-selected regions are simply copied from the preceding frame, reducing the number of computations and latency. The execution policy is trained using reinforcement learning in an online fashion without requiring ground truth annotations. Our universal framework is demonstrated on dense prediction tasks such as pedestrian detection, instance segmentation and semantic segmentation, using both state of the art (Center and Scale Predictor, MGAN, SwiftNet) and standard baseline networks (Mask-RCNN, DeepLabV3+). BlockCopy achieves significant FLOPS savings and inference speedup with minimal impact on accuracy.
****
Telling the What While Pointing to the Where: Multimodal Queries for Image Retrieval
Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, Radu Soricut; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12136-12146
Most existing image retrieval systems use text queries as a way for the user to express what they are looking for. However, fine-grained image retrieval often requires the ability to also express where in the image the content they are looking for is. The text modality can only cumbersomely express such localization preferences, whereas pointing is a more natural fit. In this paper, we propose an image retrieval setup with a new form of multimodal queries, where the user simultaneously uses both spoken natural language (the what) and mouse traces over an empty canvas (the where) to express the characteristics of the desired target image. We then describe simple modifications to an existing image retrieval model, enabling it to operate in this setup. Qualitative and quantitative experiments show that our model effectively takes this spatial guidance into account, and provides significantly more accurate retrieval results compared to text-only equi

valent systems.
****
Unsupervised Learning of Fine Structure Generation for 3D Point Clouds by 2D Pro jections Matching
Chao Chen, Zhizhong Han, Yu-Shen Liu, Matthias Zwicker; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2021, pp. 12466-12477
Learning to generate 3D point clouds without 3D supervision is an important but challenging problem. Current solutions leverage various differentiable renderers to project the generated 3D point clouds onto a 2D image plane, and train deep neural networks using the per-pixel difference with 2D ground truth images. Howe ver, these solutions are still struggling to fully recover fine structures of 3D shapes, such as thin tubes or planes. To resolve this issue, we propose an unsu pervised approach for 3D point cloud generation with fine structures. Specifical ly, we cast 3D point cloud learning as a 2D projection matching problem. Rather than using entire 2D silhouette images as a regular pixel supervision, we introd uce structure adaptive sampling to randomly sample 2D points within the silhouet tes as an irregular point supervision, which alleviates the consistency issue of sampling from different view angles. Our method pushes the neural network to ge nerate a 3D point cloud whose 2D projections match the irregular point supervisi on from different view angles. Our 2D projection matching approach enables the n eural network to learn more accurate structure information than using the per-pi xel difference, especially for fine and thin 3D structures. Our method can recov er fine 3D structures from 2D silhouette images at different resolutions, and is robust to different sampling methods and point number in irregular point superv ision. Our method outperforms others under widely used benchmarks. Our code, dat a and models are available at https://github.com/chenchao15/2D_projection_matchi ng.
****
SS-IL: Separated Softmax for Incremental Learning
Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, Taesup Moon; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202 1, pp. 844-853
We consider class incremental learning (CIL) problem, in which a learning agent continuously learns new classes from incrementally arriving training data batche s and aims to predict well on all the classes learned so far. The main challenge of the problem is the catastrophic forgetting, and for the exemplar-memory base d CIL methods, it is generally known that the forgetting is commonly caused by t he classification score bias that is injected due to the data imbalance between the new classes and the old classes (in the exemplar-memory). While several meth ods have been proposed to correct such score bias by some additional post-proces sing, e.g., score re-scaling or balanced fine-tuning, no systematic analysis on the root cause of such bias has been done. To that end, we analyze that computin g the softmax probabilities by combining the output scores for all old and new c lasses could be the main cause of the bias. Then, we propose a new CIL method, d ubbed as Separated Softmax for Incremental Learning (SS-IL), that consists of se parated softmax (SS) output layer combined with task-wise knowledge distillation (TKD) to resolve such bias. Throughout our extensive experimental results on se veral large-scale CIL benchmark datasets, we show our SS-IL achieves strong stat e-of-the-art accuracy through attaining much more balanced prediction scores acr oss old and new classes, without any additional post-processing.
****
Multiple Pairwise Ranking Networks for Personalized Video Summarization
Yassir Saquil, Da Chen, Yuan He, Chuan Li, Yong-Liang Yang; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1718-1727
In this paper, we investigate video summarization in the supervised setting. Sin ce video summarization is subjective to the preference of the end-user, the desi gn of a unique model is limited. In this work, we propose a model that provides personalized video summaries by conditioning the summarization process with pred efined categorical user labels referred to as preferences. The underlying method is based on multiple pairwise rankers (called Multi-ranker), where the rankers

are trained jointly to provide local summaries as well as a global summarization of a given video. In order to demonstrate the relevance and applications of our method in contrast with a classical global summarizer, we conduct experiments on multiple benchmark datasets, notably through a user study and comparisons with the state-of-art methods in the global video summarization task.
****

Domain-Invariant Disentangled Network for Generalizable Object Detection
Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, Jianfei Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8771-8780
We address the problem of domain generalizable object detection, which aims to learn a domain-invariant detector from multiple "seen" domains so that it can generalize well to other "unseen" domains. The generalization ability is crucial in practical scenarios especially when it is difficult to collect data. Compared to image classification, domain generalization in object detection has seldom been explored with more challenges brought by domain gaps on both image and instance levels. In this paper, we propose a novel generalizable object detection model, termed Domain-Invariant Disentangled Network (DIDN). In contrast to directly aligning multiple sources, we integrate a disentangled network into Faster R-CNN. By disentangling representations on both image and instance levels, DIDN is able to learn domain-invariant representations that are suitable for generalized object detection. Furthermore, we design a cross-level representation reconstruction to complement this two-level disentanglement so that informative object representations could be preserved. Extensive experiments are conducted on five benchmark datasets and the results demonstrate that our model achieves state-of-the-art performances on domain generalization for object detection.
****

Social NCE: Contrastive Learning of Socially-Aware Motion Representations
Yuejiang Liu, Qi Yan, Alexandre Alahi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15118-15129
Learning socially-aware motion representations is at the core of recent advances in multi-agent problems, such as human motion forecasting and robot navigation in crowds. Despite promising progress, existing representations learned with neural networks still struggle to generalize in closed-loop predictions (e.g., output colliding trajectories). This issue largely arises from the non-i.i.d. nature of sequential prediction in conjunction with ill-distributed training data. Intuitively, if the training data only comes from human behaviors in safe spaces, i.e., from "positive" examples, it is difficult for learning algorithms to capture the notion of "negative" examples like collisions. In this work, we aim to address this issue by explicitly modeling negative examples through self-supervision: (i) we introduce a social contrastive loss that regularizes the extracted motion representation by discerning the ground-truth positive events from synthetic negative ones; (ii) we construct informative negative samples based on our prior knowledge of rare but dangerous circumstances. Our method substantially reduces the collision rates of recent trajectory forecasting, behavioral cloning and reinforcement learning algorithms, outperforming state-of-the-art methods on several benchmarks. Our code is available at https://github.com/vita-epfl/social-nce.
****

The Center of Attention: Center-Keypoint Grouping via Attention for Multi-Person Pose Estimation
Guillem Brasó, Nikita Kister, Laura Leal-Taixé; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11853-11863
We introduce CenterGroup, an attention-based framework to estimate human poses from a set of identity-agnostic keypoints and person center predictions in an image. Our approach uses a transformer to obtain context-aware embeddings for all detected keypoints and centers and then applies multi-head attention to directly group joints into their corresponding person centers. While most bottom-up methods rely on non-learnable clustering at inference, CenterGroup uses a fully differentiable attention mechanism that we train end-to-end together with our keypoin

t detector. As a result, our method obtains state-of-the-art performance with up to 2.5x faster inference time than competing bottom-up methods.

****

FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters

Yuwei Cheng, Jiannan Zhu, Mengxin Jiang, Jie Fu, Changsong Pang, Peidong Wang, Kris Sankaran, Olawale Onabola, Yimin Liu, Dianbo Liu, Yoshua Bengio; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10953-10962

Marine debris is severely threatening the marine lives and causing sustained pollution to the whole ecosystem. To prevent the wastes from getting into the ocean, it is helpful to clean up the floating wastes in inland waters using the autonomous cleaning devices like unmanned surface vehicles. The cleaning efficiency relies on a high-accurate and robust object detection system. However, the small size of the target, the strong light reflection over water surface, and the reflection of other objects on bank-side all bring challenges to the vision-based object detection system. To promote the practical application for autonomous floating wastes cleaning, we present FloW, the first dataset for floating waste detection in inland water areas. The dataset consists of an image sub-dataset FloW-Img and a multimodal sub-dataset FloW-RI which contains synchronized millimeter-wave radar data and images. Accurate annotations for images and radar data are provided, supporting floating waste detection strategies based on images, radar data, and the fusion of two sensors. We perform several baseline experiments on our dataset, including vision-based and radar-based detection methods. The results show that, the detection accuracy is relatively low and floating waste detection still remains a challenging task.

****

Robust Trust Region for Weakly Supervised Segmentation

Dmitrii Marin, Yuri Boykov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6608-6618

Acquisition of training data for the standard semantic segmentation is expensive if requiring that each pixel is labeled. Yet, current methods significantly deteriorate in weakly supervised settings, e.g. where a fraction of pixels is labeled or when only image-level tags are available. It has been shown that regularized losses---originally developed for unsupervised low-level segmentation and representing geometric priors on pixel labels---can considerably improve the quality of weakly supervised training. However, many common priors require optimization stronger than gradient descent. Thus, such regularizers have limited applicability in deep learning. We propose a new robust trust region approach for regularized losses improving the state-of-the-art results. Our approach can be seen as a higher-order generalization of the classic chain rule. It allows neural network optimization to use strong low-level solvers for the corresponding regularizers, including discrete ones.

****

imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose

Thiemo Alldieck, Hongyi Xu, Cristian Sminchisescu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5461-5470

We present imGHUM, the first holistic generative model of 3D human shape and articulated pose, represented as a signed distance function. In contrast to prior work, we model the full human body implicitly as a function zero-level-set and without the use of an explicit template mesh. We propose a novel network architecture and a learning paradigm, which make it possible to learn a detailed implicit generative model of human pose, shape, and semantics, on par with state-of-the-art mesh-based models. Our model features desired detail for human models, such as articulated pose including hand motion and facial expressions, a broad spectrum of shape variations, and can be queried at arbitrary resolutions and spatial locations. Additionally, our model has attached spatial semantics making it straightforward to establish correspondences between different shape instances, thus enabling applications that are difficult to tackle using classical implicit representations. In extensive experiments, we demonstrate the model accuracy and its applicability to current research problems.

****
Dynamic High-Pass Filtering and Multi-Spectral Attention for Image Super-Resolution
Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, Hanspeter Pfister; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4288-4297
Deep convolutional neural networks (CNNs) have pushed forward the frontier of super-resolution (SR) research. However, current CNN models exhibit a major flaw: they are biased towards learning low-frequency signals. This bias becomes more problematic for the image SR task which targets reconstructing all fine details and image textures. To tackle this challenge, we propose to improve the learning of high-frequency features both locally and globally and introduce two novel architectural units to existing SR models. Specifically, we propose a dynamic high-pass filtering (HPF) module that locally applies adaptive filter weights for each spatial location and channel group to preserve high-frequency signals. We also propose a matrix multi-spectral channel attention (MMCA) module that predicts the attention map of features decomposed in the frequency domain. This module operates in a global context to adaptively recalibrate feature responses at different frequencies. Extensive qualitative and quantitative results demonstrate that our proposed modules achieve better accuracy and visual improvements against state-of-the-art methods on several benchmark datasets.
****
Self-Knowledge Distillation With Progressive Refinement of Targets
Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, Sangheum Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6567-6576
The generalization capability of deep neural networks has been substantially improved by applying a wide spectrum of regularization methods, e.g., restricting function space, injecting randomness during training, augmenting data, etc. In this work, we propose a simple yet effective regularization method named progressive self-knowledge distillation (PS-KD), which progressively distills a model's own knowledge to soften hard targets (i.e., one-hot vectors) during training. Hence, it can be interpreted within a framework of knowledge distillation as a student becomes a teacher itself. Specifically, targets are adjusted adaptively by combining the ground-truth and past predictions from the model itself. We show that PS-KD provides an effect of hard example mining by rescaling gradients according to difficulty in classifying examples. The proposed method is applicable to any supervised learning tasks with hard targets and can be easily combined with existing regularization methods to further enhance the generalization performance. Furthermore, it is confirmed that PS-KD achieves not only better accuracy, but also provides high quality of confidence estimates in terms of calibration as well as ordinal ranking. Extensive experimental results on three different tasks, image classification, object detection, and machine translation, demonstrate that our method consistently improves the performance of the state-of-the-art baselines.
****
Towards Flexible Blind JPEG Artifacts Removal
Jiaxi Jiang, Kai Zhang, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4997-5006
Training a single deep blind model to handle different quality factors for JPEG image artifacts removal has been attracting considerable attention due to its convenience for practical usage. However, existing deep blind methods usually directly reconstruct the image without predicting the quality factor, thus lacking the flexibility to control the output as the non-blind methods. To remedy this problem, in this paper, we propose a flexible blind convolutional neural network, namely FBCNN, that can predict the adjustable quality factor to control the trade-off between artifacts removal and details preservation. Specifically, FBCNN decouples the quality factor from the JPEG image via a decoupler module and then embeds the predicted quality factor into the subsequent reconstructor module through a quality factor attention block for flexible control. Besides, we find existing methods are prone to fail on non-aligned double JPEG images even with only

a one-pixel shift, and we thus propose a double JPEG degradation model to augmen t the training data. Extensive experiments on single JPEG images, more general d ouble JPEG images, and real-world JPEG images demonstrate that our proposed FBCN N achieves favorable performance against state-of-the-art methods in terms of bo th quantitative metrics and visual quality.
****

Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Rec ognition
Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, Weiming Hu; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13359-13368
Graph convolutional networks (GCNs) have been widely used and achieved remarkabl e results in skeleton-based action recognition. In GCNs, graph topology dominate s feature aggregation and therefore is the key to extracting representative feat ures. In this work, we propose a novel Channel-wise Topology Refinement Graph Co nvolution (CTR-GC) to dynamically learn different topologies and effectively agg regate joint features in different channels for skeleton-based action recognitio n. The proposed CTR-GC models channel-wise topologies through learning a shared topology as a generic prior for all channels and refining it with channel-specif ic correlations for each channel. Our refinement method introduces few extra par ameters and significantly reduces the difficulty of modeling channel-wise topolo gies. Furthermore, via reformulating graph convolutions into a unified form, we find that CTR-GC relaxes strict constraints of graph convolutions, leading to st ronger representation capability. Combining CTR-GC with temporal modeling module s, we develop a powerful graph convolutional network named CTR-GCN which notably outperforms state-of-the-art methods on the NTU RGB+D, NTU RGB+D 120, and NW-UC LA datasets.
****

VSAC: Efficient and Accurate Estimator for H and F
Maksym Ivashechkin, Daniel Barath, Ji■í Matas; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 15243-15252
We present VSAC, a RANSAC-type robust estimator with a number of novelties. It b enefits from the introduction of the concept of independent inliers that improve s significantly the efficacy of the dominant plane handling and also allows near error-free rejection of incorrect models, without false positives. The local op timization process and its application is improved so that it is run on average only once. Further technical improvements include adaptive sequential hypothesis verification and efficient model estimation via Gaussian elimination. Experimen ts on four standard datasets show that VSAC is significantly faster than all its predecessors and runs on average in 1-2 ms, on a CPU. It is two orders of magni tude faster and yet as precise as MAGSAC++, the currently most accurate estimato r of two-view geometry. In the repeated runs on EVD, HPatches, PhotoTourism, and Kusvod2 datasets, it never failed.
****

HPNet: Deep Primitive Segmentation Using Hybrid Representations
Siming Yan, Zhenpei Yang, Chongyang Ma, Haibin Huang, Etienne Vouga, Qixing Huan g; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2753-2762
This paper introduces HPNet, a novel deep-learning approach for segmenting a 3D shape represented as a point cloud into primitive patches. The key to deep primi tive segmentation is learning a feature representation that can separate points of different primitives. Unlike utilizing a single feature representation, HPNet leverages hybrid representations that combine one learned semantic descriptor, two spectral descriptors derived from predicted geometric parameters, as well as an adjacency matrix that encodes sharp edges. Moreover, instead of merely conca tenating the descriptors, HPNet optimally combines hybrid representations by lea rning combination weights. This weighting module builds on the entropy of input features. The output primitive segmentation is obtained from a mean-shift cluste ring module. Experimental results on benchmark datasets ANSI and ABCParts show t hat HPNet leads to significant performance gains from baseline approaches.

****
Fusion Moves for Graph Matching
Lisa Hutschenreiter, Stefan Haller, Lorenz Feineis, Carsten Rother, Dagmar Kainmüller, Bogdan Savchynskyy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6270-6279
We contribute to approximate algorithms for the quadratic assignment problem also known as graph matching. Inspired by the success of the fusion moves technique developed for multilabel discrete Markov random fields, we investigate its applicability to graph matching. In particular, we show how fusion moves can be efficiently combined with the dedicated state-of-the-art dual methods that have recently shown superior results in computer vision and bio-imaging applications. As our empirical evaluation on a wide variety of graph matching datasets suggests, fusion moves significantly improve performance of these methods in terms of speed and quality of the obtained solutions. Our method sets a new state-of-the-art with a notable margin with respect to its competitors.
****
Universal-Prototype Enhancing for Few-Shot Object Detection
Aming Wu, Yahong Han, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9567-9576
Few-shot object detection (FSOD) aims to strengthen the performance of novel object detection with few labeled samples. To alleviate the constraint of few samples, enhancing the generalization ability of learned features for novel objects plays a key role. Thus, the feature learning process of FSOD should focus more on intrinsical object characteristics, which are invariant under different visual changes and therefore are helpful for feature generalization. Unlike previous attempts of the meta-learning paradigm, in this paper, we explore how to enhance object features with intrinsical characteristics that are universal across different object categories. We propose a new prototype, namely universal prototype, that is learned from all object categories. Besides the advantage of characterizing invariant characteristics, the universal prototypes alleviate the impact of unbalanced object categories. After enhancing object features with the universal prototypes, we impose a consistency loss to maximize the agreement between the enhanced features and the original ones, which is beneficial for learning invariant object characteristics. Thus, we develop a new framework of few-shot object detection with universal prototypes (FSOD ^ up) that owns the merit of feature generalization towards novel objects. Experimental results on PASCAL VOC and MS COCO show the effectiveness of FSOD ^ up . Particularly, for the 1-shot case of VOC Split2, FSOD ^ up outperforms the baseline by 6.8% in terms of mAP.
****
I2UV-HandNet: Image-to-UV Prediction Network for Accurate and High-Fidelity 3D Hand Mesh Modeling
Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, Yong Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12929-12938
Reconstructing a high-precision and high-fidelity 3D human hand from a color image plays a central role in replicating a realistic virtual hand in human-computer interaction and virtual reality applications. Current methods are lacking in accuracy and fidelity due to various hand poses and severe occlusions. In this study, we propose an I2UV-HandNet model for accurate hand pose and shape estimation as well as 3D hand super-resolution reconstruction. Specifically, we present the first UV-based 3D hand shape representation. To recover a 3D hand mesh from an RGB image, we design an AffineNet to predict a UV position map from the input in an image-to-image translation fashion. To obtain a higher fidelity shape, we exploit an additional SRNet to transform the low-resolution UV map outputted by AffineNet into a high-resolution one. For the first time, we demonstrate the characterization capability of the UV-based hand shape representation. Our experiments show that the proposed method achieves state-of-the-art performance on several challenging benchmarks.
****
Fast Video Moment Retrieval

Junyu Gao, Changsheng Xu; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2021, pp. 1523-1532
This paper targets at fast video moment retrieval (fast VMR), aiming to localize
 the target moment efficiently and accurately as queried by a given natural lang
uage sentence. We argue that most existing VMR approaches can be divided into th
ree modules namely video encoder, text encoder, and cross-modal interaction modu
le, where the last module is the test-time computational bottleneck. To tackle t
his issue, we replace the cross-modal interaction module with a cross-modal comm
on space, in which moment-query alignment is learned and efficient moment search
 can be performed. For the sake of robustness in the learned space, we propose a
 fine-grained semantic distillation framework to transfer knowledge from additio
nal semantic structures. Specifically, we build a semantic role tree that decomp
oses a query sentence into different phrases (subtrees). A hierarchical semantic
-guided attention module is designed to perform message propagation across the w
hole tree and yield discriminative features. Finally, the important and discrimi
native semantics are transferred to the common space by a matching-score distill
ation process. Extensive experimental results on three popular VMR benchmarks de
monstrate that our proposed method enjoys the merits of high speed and significa
nt performance.
****

Self-Supervised Geometric Features Discovery via Interpretable Attention for Veh
icle Re-Identification and Beyond
Ming Li, Xinming Huang, Ziming Zhang; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2021, pp. 194-204
To learn distinguishable patterns, most of recent works in vehicle re-identifica
tion (ReID) struggled to redevelop official benchmarks to provide various superv
isions, which requires prohibitive human labors. In this paper, we seek to achie
ve the similar goal but do not involve more human efforts. To this end, we intro
duce a novel framework, which successfully encodes both geometric local features
 and global representations to distinguish vehicle instances, optimized only by
the supervision from official ID labels. Specifically, given our insight that ob
jects in ReID share similar geometric characteristics, we propose to borrow self
-supervised representation learning to facilitate geometric features discovery.
To condense these features, we introduce an interpretable attention module, with
 the core of local maxima aggregation instead of fully automatic learning, whose
 mechanism is completely understandable and whose response map is physically rea
sonable. To the best of our knowledge, we are the first that perform self-superv
ised learning to discover geometric features. We conduct comprehensive experimen
ts on three most popular datasets for vehicle ReID, i.e., VeRi-776, CityFlow-ReI
D, and VehicleID. We report our state-of-the-art (SOTA) performances and promisi
ng visualization results. We also show the excellent scalability of our approach
 on other ReID related tasks, i.e., person ReID and multi-target multi-camera (M
TMC) vehicle tracking.
****

With a Little Help From My Friends: Nearest-Neighbor Contrastive Learning of Vis
ual Representations
Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisser
man; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2021, pp. 9588-9597
Self-supervised learning algorithms based on instance discrimination train encod
ers to be invariant to pre-defined transformations of the same instance. While m
ost methods treat different views of the same image as positives for a contrasti
ve loss, we are interested in using positives from other instances in the datase
t. Our method, Nearest-Neighbor Contrastive Learning of visual Representations (
NNCLR), samples the nearest neighbors from the dataset in the latent space, and
treats them as positives. This provides more semantic variations than pre-define
d transformations. We find that using the nearest-neighbor as positive in contra
stive losses improves performance significantly on ImageNet classification, from
 71.7% to 75.6%, outperforming previous state-of-the-art methods. On semi-superv
ised learning benchmarks we improve performance significantly when only 1% Image

Net labels are available, from 53.8% to 56.5%. On transfer learning benchmarks o ur method outperforms state-of-the-art methods (including supervised learning wi th ImageNet) on 8 out of 12 downstream datasets. Furthermore, we demonstrate emp irically that our method is less reliant on complex data augmentations. We see a relative reduction of only 2.1% ImageNet Top-1 accuracy when we train using onl y random crops.
****

Explainable Person Re-Identification With Attribute-Guided Metric Distillation
Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, Tao Mei; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 11813-11822
Despite the great progress of person re-identification (ReID) with the adoption of Convolutional Neural Networks, current ReID models are opaque and only output s a scalar distance between two persons. There are few methods providing users s emantically understandable explanations for why two persons are the same one or not. In this paper, we propose a post-hoc method, named Attribute-guided Metric Distillation (AMD), to explain existing ReID models. This is the first method to explore attributes to answer: 1) what and where the attributes make two persons different, and 2) how much each attribute contributes to the difference. In AMD , we design a pluggable interpreter network for target models to generate quanti tative contributions of attributes and visualize accurate attention maps of the most discriminative attributes. To achieve this goal, we propose a metric distil lation loss by which the interpreter learns to decompose the distance of two per sons into components of attributes with knowledge distilled from the target mode l. Moreover, we propose an attribute prior loss to make the interpreter generate attribute-guided attention maps and to eliminate biases caused by the imbalance d distribution of attributes. This loss can guide the interpreter to focus on th e exclusive and discriminative attributes rather than the large-area but common attributes of two persons. Comprehensive experiments show that the interpreter c an generate effective and intuitive explanations for varied models and generaliz e well under cross-domain settings. As a by-product, the accuracy of target mode ls can be further improved with our interpreter.
****

Motion-Focused Contrastive Learning of Video Representations
Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, Tao Mei; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2105-2 114
Motion, as the most distinct phenomenon in a video to involve the changes over t ime, has been unique and critical to the development of video representation lea rning. In this paper, we ask the question: how important is the motion particula rly for self-supervised video representation learning. To this end, we compose a duet of exploiting the motion for data augmentation and feature learning in the regime of contrastive learning. Specifically, we present a Motion-focused Contr astive Learning (MCL) method that regards such duet as the foundation. On one ha nd, MCL capitalizes on optical flow of each frame in a video to temporally and s patially sample the tubelets (i.e., sequences of associated frame patches across time) as data augmentations. On the other hand, MCL further aligns gradient map s of the convolutional layers to optical flow maps from spatial, temporal and sp atio-temporal perspectives, in order to ground motion information in feature lea rning. Extensive experiments conducted on R(2+1)D backbone demonstrate the effec tiveness of our MCL. On UCF101, the linear classifier trained on the representat ions learnt by MCL achieves 81.91% top-1 accuracy, outperforming ImageNet superv ised pre-training by 6.78%. On Kinetics-400, MCL achieves 66.62% top-1 accuracy under the linear protocol.
****

Motion Guided Region Message Passing for Video Captioning
Shaoxiang Chen, Yu-Gang Jiang; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 1543-1552
Video captioning is an important vision task and has been intensively studied in the computer vision community. Existing methods that utilize the fine-grained s

patial information have achieved significant improvements, however, they either rely on costly external object detectors or do not sufficiently model the spatial/temporal relations. In this paper, we aim at designing a spatial information extraction and aggregation method for video captioning without the need of external object detectors. For this purpose, we propose a Recurrent Region Attention module to better extract diverse spatial features, and by employing Motion-Guided Cross-frame Message Passing, our model is aware of the temporal structure and able to establish high-order relations among the diverse regions across frames. They jointly encourage information communication and produce compact and powerful video representations. Furthermore, an Adjusted Temporal Graph Decoder is proposed to flexibly update video features and model high-order temporal relations during decoding. Experimental results on three benchmark datasets: MSVD, MSR-VTT, and VATEX demonstrate that our proposed method can outperform state-of-the-art methods.

****

Learning Causal Representation for Training Cross-Domain Pose Estimator via Generative Interventions

Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, Weidong Geng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11270-11280

3D pose estimation has attracted increasing attention with the availability of high-quality benchmark datasets. However, prior works show that deep learning models tend to learn spurious correlations, which fail to generalize beyond the specific dataset they are trained on. In this work, we take a step towards training robust models for cross-domain pose estimation task, which brings together ideas from causal representation learning and generative adversarial networks. Specifically, this paper introduces a novel framework for causal representation learning which explicitly exploits the causal structure of the task. We consider changing domain as interventions on images under the data-generation process and steer the generative model to produce counterfactual features. This help the model learn transferable and causal relations across different domains. Our framework is able to learn with various types of unlabeled datasets. We demonstrate the efficacy of our proposed method on both human and hand pose estimation task. The experiment results show the proposed approach achieves state-of-the-art performance on most datasets for both domain adaptation and domain generalization settings.

****

Super-Resolving Cross-Domain Face Miniatures by Peeking at One-Shot Exemplar

Peike Li, Xin Yu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4469-4479

Conventional face super-resolution methods usually assume testing low-resolution (LR) images lie in the same domain as the training ones. Due to different lighting conditions and imaging hardware, domain gaps between training and testing images inevitably occur in many real-world scenarios. Neglecting those domain gaps would lead to inferior face super-resolution (FSR) performance. However, how to transfer a trained FSR model to a target domain efficiently and effectively has not been investigated. To tackle this problem, we develop a Domain-Aware Pyramid-based Face Super-Resolution network, named DAP-FSR network. Our DAP-FSR is the first attempt to super-resolve LR faces from a target domain by exploiting only a pair of high-resolution (HR) and LR exemplar in the target domain. To be specific, our DAP-FSR firstly employs its encoder to extract the multi-scale latent representations of the input LR face. Considering only one target domain example is available, we propose to augment the target domain data by mixing the latent representations of the target domain face and source domain ones and then feed the mixed representations to the decoder of our DAP-FSR. The decoder will generate new face images resembling the target domain image style. The generated HR faces in turn are used to optimize our decoder to reduce the domain gap. By iteratively updating the latent representations and our decoder, our DAP-FSR will be adapted to the target domain, thus achieving authentic and high-quality upsampled HR faces. Extensive experiments on three benchmarks validate the effectiveness

and superior performance of our DAP-FSR compared to the state-of-the-art methods
.
****
Webly Supervised Fine-Grained Recognition: Benchmark Datasets and an Approach
Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jia
n Zhang, Heng Tao Shen; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2021, pp. 10602-10611
Learning from the web can ease the extreme dependence of deep learning on large-
scale manually labeled datasets. Especially for fine-grained recognition, which
targets at distinguishing subordinate categories, it will significantly reduce t
he labeling costs by leveraging free web data. Despite its significant practical
 and research value, the webly supervised fine-grained recognition problem is no
t extensively studied in the computer vision community, largely due to the lack
of high-quality datasets. To fill this gap, in this paper we construct two new b
enchmark webly supervised fine-grained datasets, termed WebFG-496 and WebiNat-50
89, respectively. In concretely, WebFG-496 consists of three sub-datasets contai
ning a total of 53,339 web training images with 200 species of birds (Web-bird),
 100 types of aircrafts (Web-aircraft), and 196 models of cars (Web-car). For We
biNat-5089, it contains 5089 sub-categories and more than 1.1 million web traini
ng images, which is the largest webly supervised fine-grained dataset ever. As a
 minor contribution, we also propose a novel webly supervised method (termed "Pe
er-learning") for benchmarking these datasets. Comprehensive experimental result
s and analyses on two new benchmark datasets demonstrate that the proposed metho
d achieves superior performance over the competing baseline models and states-of
-the-art.
****
Towards Interpretable Deep Networks for Monocular Depth Estimation
Zunzhi You, Yi-Hsuan Tsai, Wei-Chen Chiu, Guanbin Li; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2021, pp. 12879-12888
Deep networks for Monocular Depth Estimation (MDE) have achieved promising perfo
rmance recently and it is of great importance to further understand the interpre
tability of these networks. Existing methods attempt to provide post-hoc explana
tions by investigating visual cues, which may not explore the internal represent
ations learned by deep networks. In this paper, we find that some hidden units o
f the network are selective to certain ranges of depth, and thus such behavior c
an be served as a way to interpret the internal representations. Based on our ob
servations, we quantify the interpretability of a deep MDE network by the depth
selectivity of its hidden units. Moreover, we then propose a method to train int
erpretable MDE deep networks without changing their original architectures, by a
ssigning a depth range for each unit to select. Experimental results demonstrate
 that our method is able to enhance the interpretability of deep MDE networks by
 largely improving the depth selectivity of their units, while not harming or ev
en improving the depth estimation accuracy. We further provide comprehensive ana
lysis to show the reliability of selective units, the applicability of our metho
d on different models and layers, and a demonstration on monocular depth complet
ion. We further provide comprehensive analysis to show the reliability of select
ive units, the applicability of our method on different layers, models, and data
sets, and a demonstration on analysis of model error.
****
Instance Segmentation in 3D Scenes Using Semantic Superpoint Tree Networks
Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, Kui Jia; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2783-2792
Instance segmentation in 3D scenes is fundamental in many applications of scene
understanding. It is yet challenging due to the compound factors of data irregul
arity and uncertainty in the numbers of instances. State-of-the-art methods larg
ely rely on a general pipeline that first learns point-wise features discriminat
ive at semantic and instance levels, followed by a separate step of point groupi
ng for proposing object instances. While promising, they have the shortcomings t
hat (1) the second step is not supervised by the main objective of instance segm
entation, and (2) their point-wise feature learning and grouping are less effect

ive to deal with data irregularities, possibly resulting in fragmented segmentat ions. To address these issues, we propose in this work an end-to-end solution of Semantic Superpoint Tree Network (SSTNet) for proposing object instances from s cene points. Key in SSTNet is an intermediate, semantic superpoint tree (SST), w hich is constructed based on the learned semantic features of superpoints, and w hich will be traversed and split at intermediate tree nodes for proposals of obj ect instances. We also design in SSTNet a refinement module, termed CliqueNet, t o prune superpoints that may be wrongly grouped into instance proposals. Experim ents on the benchmarks of ScanNet and S3DIS show the efficacy of our proposed me thod. At the time of submission, SSTNet ranks top on the ScanNet (V2) leaderboar d, with 2% higher of mAP than the second best method.
****

Exploring Cross-Image Pixel Contrast for Semantic Segmentation
Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, Luc Van Gool ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2021, pp. 7303-7313
Current semantic segmentation methods focus only on mining "local" context, i.e. , dependencies between pixels within individual images, by context-aggregation m odules (e.g., dilated convolution, neural attention) or structure-aware optimiza tion criteria (e.g., IoU-like loss). However, they ignore "global" context of th e training data, i.e., rich semantic relations between pixels across different i mages. Inspired by recent advance in unsupervised contrastive representation lea rning, we propose a pixel-wise contrastive algorithm for semantic segmentation i n the fully supervised setting. The core idea is to enforce pixel embeddings bel onging to a same semantic class to be more similar than embeddings from differen t classes. It raises a pixel-wise metric learning paradigm for semantic segmenta tion, by explicitly exploring the structures of labeled pixels, which were rarel y explored before. Our method can be effortlessly incorporated into existing seg mentation frameworks without extra overhead during testing. We experimentally sh ow that, with famous segmentation models (i.e., DeepLabV3, HRNet, OCR) and backb ones (i.e., ResNet, HRNet), our method brings performance improvements across di verse datasets (i.e., Cityscapes, PASCAL-Context, COCO-Stuff, CamVid). We expect this work will encourage our community to rethink the current de facto training paradigm in semantic segmentation.
****

Geometric Granularity Aware Pixel-To-Mesh
Yue Shi, Bingbing Ni, Jinxian Liu, Dingyi Rong, Ye Qian, Wenjun Zhang; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13097-13106
Pixel-to-mesh has wide applications, especially in virtual or augmented reality, animation and game industry. However, existing mesh reconstruction models perfo rm unsatisfactorily in local geometry details due to ignoring mesh topology info rmation during learning. Besides, most methods are constrained by the initial te mplate, which cannot reconstruct meshes of various genus. In this work, we propo se a geometric granularity-aware pixel-to-mesh framework with a fidelity-selecti on-and-guarantee strategy, which explicitly addresses both challenges. First, a geometry structure extractor is proposed for detecting local high structured par ts and capturing local spatial feature. Second, we apply it to facilitate pixel- to-mesh mapping and resolve coarse details problem caused by the neglect of stru ctural information in previous practices. Finally, a mesh edit module is propose d to encourage non-zero genus topology to emergence by fine-grained topology mod ification and a patching algorithm is introduced to repair the non-closed bounda ries. Extensive experimental results, both quantitatively and visually have demo nstrated the high reconstruction fidelity achieved by the proposed framework.
****

Pixel Difference Networks for Efficient Edge Detection
Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, Li Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5117-5127
Recently, deep Convolutional Neural Networks (CNNs) can achieve human-level perf

ormance in edge detection with the rich and abstract edge representation capacities. However, the high performance of CNN based edge detection is achieved with a large pretrained CNN backbone, which is memory and energy consuming. In addition, it is surprising that the previous wisdom from the traditional edge detectors, such as Canny, Sobel, and LBP are rarely investigated in the rapid-developing deep learning era. To address these issues, we propose a simple, lightweight yet effective architecture named Pixel Difference Network (PiDiNet) for efficient edge detection. Extensive experiments on BSDS500, NYUD, and Multicue are provided to demonstrate its effectiveness, and its high training and inference efficiency. Surprisingly, when training from scratch with only the BSDS500 and VOC datasets, PiDiNet can surpass the recorded result of human perception (0.807 vs. 0.803 in ODS F-measure) on the BSDS500 dataset with 100 FPS and less than 1M parameters. A faster version of PiDiNet with less than 0.1M parameters can still achieve comparable performance among state of the arts with 200 FPS. Results on the NYUD and Multicue datasets show similar observations. The codes are available at https://github.com/zhuoinoulu/pidinet.

****

Towards Understanding the Generative Capability of Adversarially Robust Classifiers

Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, Yaowu Chen, Zhenguo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7728-7737

Recently, some works found an interesting phenomenon that adversarially robust classifiers can generate good images comparable to generative models. We investigate this phenomenon from an energy perspective and provide a novel explanation. We reformulate adversarial example generation, adversarial training, and image generation in terms of an energy function. We find that adversarial training contributes to obtaining an energy function that is flat and has low energy around the real data, which is the key for generative capability. Based on our new understanding, we further propose a better adversarial training method, Joint Energy Adversarial Training (JEAT), which can generate high-quality images and achieve new state-of-the-art robustness under a wide range of attacks. The Inception Score of the images (CIFAR-10) generated by JEAT is 8.80, much better than original robust classifiers (7.50). In particular, we achieve new state-of-the-art robustness on CIFAR-10 (from 57.20% to 62.04%) and CIFAR-100 (from 30.03% to 30.18%) without extra training data.

****

Learning Efficient Photometric Feature Transform for Multi-View Stereo

Kaizhang Kang, Cihui Xie, Ruisheng Zhu, Xiaohe Ma, Ping Tan, Hongzhi Wu, Kun Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5956-5965

We present a novel framework to learn to convert the per-pixel photometric information at each view into spatially distinctive and view-invariant low-level features, which can be plugged into existing multi-view stereo pipeline for enhanced 3D reconstruction. Both the illumination conditions during acquisition and the subsequent per-pixel feature transform can be jointly optimized in a differentiable fashion. Our framework automatically adapts to and makes efficient use of the geometric information available in different forms of input data. High-quality 3D reconstructions of a variety of challenging objects are demonstrated on the data captured with an illumination multiplexing device, as well as a point light. Our results compare favorably with state-of-the-art techniques.

****

NEAT: Neural Attention Fields for End-to-End Autonomous Driving

Kashyap Chitta, Aditya Prakash, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15793-15803

Efficient reasoning about the semantic, spatial, and temporal structure of a scene is a crucial prerequisite for autonomous driving. We present NEural ATtention fields (NEAT), a novel representation that enables such reasoning for end-to-end imitation learning models. NEAT is a continuous function which maps locations in Bird's Eye View (BEV) scene coordinates to waypoints and semantics, using int

ermediate attention maps to iteratively compress high-dimensional 2D image features into a compact representation. This allows our model to selectively attend to relevant regions in the input while ignoring information irrelevant to the driving task, effectively associating the images with the BEV representation. In a new evaluation setting involving adverse environmental conditions and challenging scenarios, NEAT outperforms several strong baselines and achieves driving scores on par with the privileged CARLA expert used to generate its training data. Furthermore, visualizing the attention maps for models with NEAT intermediate representations provides improved interpretability.
****
Modulated Periodic Activations for Generalizable Local Functional Representations
Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, Manmohan Chandraker; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14214-14223
Multi-Layer Perceptrons (MLPs) make powerful functional representations for sampling and reconstruction problems involving low-dimensional signals like images,shapes and light fields. Recent works have significantly improved their ability to represent high-frequency content by using periodic activations or positional encodings. This often came at the expense of generalization: modern methods are typically optimized for a single signal. We present a new representation that generalizes to multiple instances and achieves state-of-the-art fidelity. We use a dual-MLP architecture to encode the signals. A synthesis network creates a functional mapping from a low-dimensional input(e.g. pixel-position) to the output domain (e.g. RGB color).A modulation network maps a latent code corresponding to the target signal to parameters that modulate the periodic activations of the synthesis network. We also propose a local-functional representation which enables generalization. The signal's domain is partitioned into a regular grid,with each tile represented by a latent code. At test time, the signal is encoded with high-fidelity by inferring (or directly optimizing) the latent code-book. Our approach produces generalizable functional representations of images, videos and shapes, and achieves higher reconstruction quality than prior works that are optimized for a single signal.
****
Neural Architecture Search for Joint Human Parsing and Pose Estimation
Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, Wu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11385-11394
Human parsing and pose estimation are crucial for the understanding of human behaviors. Since these tasks are closely related, employing one unified model to perform two tasks simultaneously allows them to benefit from each other. However, since human parsing is a pixel-wise classification process while pose estimation is usually a regression task, it is non-trivial to extract discriminative features for both tasks while modeling their correlation in the joint learning fashion. Recent studies have shown that Neural Architecture Search (NAS) has the ability to allocate efficient feature connections for specific tasks automatically. With the spirit of NAS, we propose to search for an efficient network architecture (NPPNet) to tackle two tasks at the same time. On the one hand, to extract task-specific features for the two tasks and lay the foundation for the further searching of feature interaction, we propose to search their encoder-decoder architectures, respectively. On the other hand, to ensure two tasks fully communicate with each other, we propose to embed NAS units in both multi-scale feature interaction and high-level feature fusion to establish optimal connections between two tasks. Experimental results on both parsing and pose estimation benchmark datasets have demonstrated that the searched model achieves state-of-the-art performances on both tasks.
****
Fast Light-Field Disparity Estimation With Multi-Disparity-Scale Cost Aggregation
Zhicong Huang, Xuemei Hu, Zhou Xue, Weizhu Xu, Tao Yue; Proceedings of the IEEE/

CVF International Conference on Computer Vision (ICCV), 2021, pp. 6320-6329
Light field images contain both angular and spatial information of captured ligh t rays. The rich information of light fields enables straightforward disparity r ecovery capability but demands high computational cost as well. In this paper, w e design a lightweight disparity estimation model with physical-based multi-disp arity-scale cost volume aggregation for fast disparity estimation. By introducin g a sub-network of edge guidance, we significantly improve the recovery of geome tric details near edges and improve the overall performance. We test the propose d model extensively on both synthetic and real-captured datasets, which provide both densely and sparsely sampled light fields. Finally, we significantly reduce computation cost and GPU memory consumption, while achieving comparable perform ance with state-of-the-art disparity estimation methods for light fields. Our so urce code is available at https://github.com/zcong17huang/FastLFnet.
****

SemIE: Semantically-Aware Image Extrapolation
Bholeshwar Khurana, Soumya Ranjan Dash, Abhishek Bhatia, Aniruddha Mahapatra, Hr ituraj Singh, Kuldeep Kulkarni; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 14900-14909
We propose a semantically-aware novel paradigm to perform image extrapolation th at enables the addition of new object instances. All previous methods are limite d in their capability of extrapolation to merely extending the already existing objects in the image. However, our proposed approach focuses not only on (i) ext ending the already present objects but also on (ii) adding new objects in the ex tended region based on the context. To this end, for a given image, we first obt ain an object segmentation map using a state-of-the-art semantic segmentation me thod. The, thus, obtained segmentation map is fed into a network to compute the extrapolated semantic segmentation and the corresponding panoptic segmentation m aps. The input image and the obtained segmentation maps are further utilized to generate the final extrapolated image. We conduct experiments on Cityscapes and ADE20K bedroom datasets and show that our method outperforms all baselines in te rms of FID, and similarity object co-occurrence statistics.
****

Transformer-Based Dual Relation Graph for Multi-Label Image Recognition
Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, Jia Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16 3-172
The simultaneous recognition of multiple objects in one image remains a challeng ing task, spanning multiple events in the recognition field such as various obje ct scales, inconsistent appearances, and confused inter-class relationships. Rec ent research efforts mainly resort to the statistic label co-occurrences and lin guistic word embedding to enhance the unclear semantics. Different from these re searches, in this paper, we propose a novel Transformer-based Dual Relation lear ning framework, constructing complementary relationships by exploring two aspect s of correlation, i.e., structural relation graph and semantic relation graph. T he structural relation graph aims to capture long-range correlations from object context, by developing a cross-scale transformer-based architecture. The semant ic graph dynamically models the semantic meanings of image objects with explicit semantic-aware constraints. In addition, we also incorporate the learnt structu ral relationship into the semantic graph, constructing a joint relation graph fo r robust representations. With the collaborative learning of these two effective relation graphs, our approach achieves new state-of-the-art on two popular mult i-label recognition benchmarks, i.e. MS-COCO and VOC 2007 dataset.
****

Self-Supervised Transfer Learning for Hand Mesh Recovery From Binocular Images
Zheng Chen, Sihan Wang, Yi Sun, Xiaohong Ma; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 11626-11634
Traditional methods for RGB hand mesh recovery usually need to train a separate model for each dataset with the corresponding ground truth and are hardly adapte d to new scenarios without the ground truth for supervision. To address the prob lem, we propose a self-supervised framework for hand mesh estimation, where we p

re-learn hand priors from existing hand datasets and transfer the priors to new scenarios without any landmark annotations. The proposed approach takes binocular images as input and mainly relies on left-right consistency constraints including appearance consensus and shape consistency to train the model to estimate the hand mesh in new scenarios. We conduct experiments on the widely used stereo hand dataset, and the experimental results verify that our model can get comparable performance compared with state-of-the-art methods even without the corresponding landmark annotations. To further evaluate our model, we collect a large real binocular dataset. The experimental results on the collected real dataset also verify the effectiveness of our model qualitatively.
****

Faster Multi-Object Segmentation Using Parallel Quadratic Pseudo-Boolean Optimization
Niels Jeppesen, Patrick M. Jensen, Anders N. Christensen, Anders B. Dahl, Vedrana A. Dahl; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6260-6269
We introduce a parallel version of the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm for solving binary optimization tasks, such as image segmentation. The original QPBO implementation by Kolmogorov and Rother relies on the Boykov-Kolmogorov (BK) maxflow/mincut algorithm and performs well for many image analysis tasks. However, the serial nature of their QPBO algorithm results in poor utilization of modern hardware. By redesigning the QPBO algorithm to work with parallel maxflow/mincut algorithms, we significantly reduce solve time of large optimization tasks. We compare our parallel QPBO implementation to other state-of-the-art solvers and benchmark them on two large segmentation tasks and a substantial set of small segmentation tasks. The results show that our parallel QPBO algorithm is over 20 times faster than the serial QPBO algorithm on the large tasks and over three times faster for the majority of the small tasks. Although we focus on image segmentation, our algorithm is generic and can be used for any QPBO problem. Our implementation and experimental results are available at DOI: 10.5281/zenodo.5201620
****

Partial Off-Policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning
Jiahe Shi, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2187-2196
Human-oriented image captioning with both high diversity and accuracy is a challenging task in vision+language modeling. The reinforcement learning (RL) based frameworks promote the accuracy of image captioning, yet seriously hurt the diversity. In contrast, other methods based on variational auto-encoder (VAE) or generative adversarial network (GAN) can produce diverse yet less accurate captions. In this work, we devote our attention to promote the diversity of RL-based image captioning. To be specific, we devise a partial off-policy learning scheme to balance accuracy and diversity. First, we keep the model exposed to varied candidate captions by sampling from the initial state before RL launched. Second, a novel criterion named max-CIDEr is proposed to serve as the reward for promoting diversity. We combine the above-mentioned off-policy strategy with the on-policy one to moderate the exploration effect, further balancing the diversity and accuracy for human-like image captioning. Experiments show that our method locates the closest to human performance in the diversity-accuracy space, and achieves the highest Pearson correlation as 0.337 with human performance.
****

Prior to Segment: Foreground Cues for Weakly Annotated Classes in Partially Supervised Instance Segmentation
David Biertimpel, Sindi Shkodrani, Anil S. Baslamisli, Nóra Baka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2824-2833
Instance segmentation methods require large datasets with expensive and thus limited instance-level mask labels. Partially supervised instance segmentation aims to improve mask prediction with limited mask labels by utilizing the more abund

ant weak box labels. In this work, we show that a class agnostic mask head, comm only used in partially supervised instance segmentation, has difficulties learni ng a general concept of foreground for the weakly annotated classes using box su pervision only. To resolve this problem, we introduce an object mask prior (OMP) that provides the mask head with the general concept of foreground implicitly l earned by the box classification head under the supervision of all classes. This helps the class agnostic mask head to focus on the primary object in a region o f interest (RoI) and improves generalization to the weakly annotated classes. We test our approach on the COCO dataset using different splits of strongly and we akly supervised classes. Our approach significantly improves over the Mask R-CNN baseline and obtains competitive performance with the state-of-the-art, while o ffering a much simpler architecture.

****

Interpretation of Emergent Communication in Heterogeneous Collaborative Embodied Agents

Shivansh Patel, Saim Wani, Unnat Jain, Alexander G. Schwing, Svetlana Lazebnik, Manolis Savva, Angel X. Chang; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 15953-15963

Communication between embodied AI agents has received increasing attention in re cent years. Despite its use, it is still unclear whether the learned communicati on is interpretable and grounded in perception. To study the grounding of emerge nt forms of communication, we first introduce the collaborative multi-object nav igation task 'CoMON.' In this task, an 'oracle agent' has detailed environment i nformation in the form of a map. It communicates with a 'navigator agent' that p erceives the environment visually and is tasked to find a sequence of goals. To succeed at the task, effective communication is essential. CoMON hence serves as a basis to study different communication mechanisms between heterogeneous agent s, that is, agents with different capabilities and roles. We study two common co mmunication mechanisms and analyze their communication patterns through an egoce ntric and spatial lens. We show that the emergent communication can be grounded to the agent observations and the spatial structure of the 3D environment.

****

A Dark Flash Normal Camera

Zhihao Xia, Jason Lawrence, Supreeth Achar; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2021, pp. 2430-2439

Casual photography is often performed in uncontrolled lighting that can result i n low quality images and degrade the performance of downstream processing. We co nsider the problem of estimating surface normal and reflectance maps of scenes d epicting people despite these conditions by supplementing the available visible illumination with a single near infrared (NIR) light source and camera, a so-cal led "dark flash image". Our method takes as input a single color image captured under arbitrary visible lighting and a single dark flash image captured under co ntrolled front-lit NIR lighting at the same viewpoint, and computes a normal map , a diffuse albedo map, and a specular intensity map of the scene. Since ground truth normal and reflectance maps of faces are difficult to capture, we propose a novel training technique that combines information from two readily available and complementary sources: a stereo depth signal and photometric shading cues. W e evaluate our method over a range of subjects and lighting conditions and descr ibe two applications: optimizing stereo geometry and filling the shadows in an i mage.

****

Dynamic CT Reconstruction From Limited Views With Implicit Neural Representation s and Parametric Motion Fields

Albert W. Reed, Hyojin Kim, Rushil Anirudh, K. Aditya Mohan, Kyle Champley, Jing u Kang, Suren Jayasuriya; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2021, pp. 2258-2268

Reconstructing dynamic, time-varying scenes with computed tomography (4D-CT) is a challenging and ill-posed problem common to industrial and medical settings. E xisting 4D-CT reconstructions are designed for sparse sampling schemes that requ ire fast CT scanners to capture multiple, rapid revolutions around the scene in

order to generate high quality results. However, if the scene is moving too fast , then the sampling occurs along a limited view and is difficult to reconstruct due to spatiotemporal ambiguities. In this work, we design a reconstruction pipe line using implicit neural representations coupled with a novel parametric motio n field warping to perform limited view 4D-CT reconstruction of rapidly deformin g scenes. Importantly, we utilize a differentiable analysis-by-synthesis approac h to compare with captured x-ray sinogram data in a self-supervised fashion. Thu s, our resulting optimization method requires no training data to reconstruct th e scene. We demonstrate that our proposed system robustly reconstructs scenes co ntaining deformable and periodic motion and validate against state-of-the-art ba selines. Further, we demonstrate an ability to reconstruct continuous spatiotemp oral representations of our scenes and upsample them to arbitrary volumes and fr ame rates post-optimization. This research opens a new avenue for implicit neura l representations in computed tomography reconstruction in general. Code is avai lable at https://github.com/awreed/DynamicCTReconstruction.

****

Diverse Image Style Transfer via Invertible Cross-Space Mapping

Haibo Chen, Lei Zhao, Huiming Zhang, Zhizhong Wang, Zhiwen Zuo, Ailin Li, Wei Xi ng, Dongming Lu; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2021, pp. 14880-14889

Image style transfer aims to transfer the styles of artworks onto arbitrary phot ographs to create novel artistic images. Although style transfer is inherently a n underdetermined problem, existing approaches usually assume a deterministic so lution, thus failing to capture the full distribution of possible outputs. To ad dress this limitation, we propose a Diverse Image Style Transfer (DIST) framewor k which achieves significant diversity by enforcing an invertible cross-space ma pping. Specifically, the framework consists of three branches: disentanglement b ranch, inverse branch, and stylization branch. Among them, the disentanglement b ranch factorizes artworks into content space and style space; the inverse branch encourages the invertible mapping between the latent space of input noise vecto rs and the style space of generated artistic images; the stylization branch rend ers the input content image with the style of an artist. Armed with these three branches, our approach is able to synthesize significantly diverse stylized imag es without loss of quality. We conduct extensive experiments and comparisons to evaluate our approach qualitatively and quantitatively. The experimental results demonstrate the effectiveness of our method.

****

Variational Attention: Propagating Domain-Specific Knowledge for Multi-Domain Le arning in Crowd Counting

Binghui Chen, Zhaoyi Yan, Ke Li, Pengyu Li, Biao Wang, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16065-16075

In crowd counting, due to the problem of laborious labelling, it is perceived in tractability of collecting a new large-scale dataset which has plentiful images with large diversity in density, scene, etc. Thus, for learning a general model, training with data from multiple different datasets might be a remedy and be of great value. In this paper, we resort to the multi-domain joint learning and pr opose a simple but effective Domain-specific Knowledge Propagating Network (DKPN et) for unbiasedly learning the knowledge from multiple diverse data domains at the same time. It is mainly achieved by proposing the novel Variational Attentio n(VA) technique for explicitly modeling the attention distributions for differen t domains. And as an extension to VA, Intrinsic Variational Attention(InVA) is p roposed to handle the problems of over-lapped domains and sub-domains. Extensive experiments have been conducted to validate the superiority of our DKPNet over several popular datasets, including ShanghaiTech A/B, UCF-QNRF and NWPU.

****

Pri3D: Can 3D Priors Help 2D Representation Learning?

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 56 93-5702

Recent advances in 3D perception have shown impressive progress in understanding geometric structures of 3D shapes and even scenes. Inspired by these advances in geometric understanding, we aim to imbue image-based perception with representations learned under geometric constraints. We introduce an approach to learn view-invariant, geometry-aware representations for network pre-training, based on multi-view RGB-D data, that can then be effectively transferred to downstream 2D tasks. We propose to employ contrastive learning under both multi-view image constraints and image-geometry constraints to encode 3D priors into learned 2D representations. This results not only in improvement over 2D-only representation learning on the image-based tasks of semantic segmentation, instance segmentation, and object detection on real-world indoor datasets, but moreover, provides significant improvement in the low data regime. We show a significant improvement of 6.0% on semantic segmentation on full data as well as 11.9% on 20% data against our baselines on ScanNet.

****

PoGO-Net: Pose Graph Optimization With Graph Neural Networks

Xinyi Li, Haibin Ling; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5895-5905

Accurate camera pose estimation or global camera re-localization is a core component in Structure-from-Motion (SfM) and SLAM systems. Given pair-wise relative camera poses, pose-graph optimization (PGO) involves solving for an optimized set of globally-consistent absolute camera poses. In this work, we propose a novel PGO scheme fueled by graph neural networks (GNN), namely PoGO-Net, to conduct the absolute camera pose regression leveraging multiple rotation averaging (MRA). Specifically, PoGO-Net takes a noisy view-graph as the input, where the nodes and edges are designed to encode the geometric constraints and local graph consistency. Besides, we address the outlier edge removal by exploiting an implicit edge-dropping scheme where the noisy or corrupted edges are effectively filtered out with parameterized networks. Furthermore, we introduce a joint loss function embedding MRA formulation such that the robust inference is capable of achieving real-time performances even for large-scale scenes. Our proposed network is trained end-to-end on public benchmarks, outperforming state-of-the-art approaches in extensive experiments that demonstrate the efficiency and robustness of our proposed network.

****

Federated Learning for Non-IID Data via Unified Feature Learning and Optimization Objective Alignment

Lin Zhang, Yong Luo, Yan Bai, Bo Du, Ling-Yu Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4420-4428

Federated Learning (FL) aims to establish a shared model across decentralized clients under the privacy-preserving constraint. Despite certain success, it is still challenging for FL to deal with non-IID (non-independent and identical distribution) client data, which is a general scenario in real-world FL tasks. It has been demonstrated that the performance of FL will be reduced greatly under the non-IID scenario, since the discrepant data distributions will induce optimization inconsistency and feature divergence issues. Besides, naively minimizing an aggregate loss function in this scenario may have negative impacts on some clients and thus deteriorate their personal model performance. To address these issues, we propose a Unified Feature learning and Optimization objectives alignment method (FedUFO) for non-IID FL. In particular, an adversary module is proposed to reduce the divergence on feature representation among different clients, and two consensus losses are proposed to reduce the inconsistency on optimization objectives from two perspectives. Extensive experiments demonstrate that our FedUFO can outperform the state-of-the-art approaches, including the competitive one data-sharing method. Besides, FedUFO can enable more reasonable and balanced model performance among different clients.

****

Self-Supervised Video Object Segmentation by Motion Grouping

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, Weidi Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 717

7-7188
Animals have evolved highly functional visual systems to understand motion, assisting perception even under complex environments. In this paper, we work towards developing a computer vision system able to segment objects by exploiting motion cues, i.e. motion segmentation. To achieve this, we introduce a simple variant of the Transformer to segment optical flow frames into primary objects and the background, which can be trained in a self-supervised manner, i.e. without using any manual annotations. Despite using only optical flow, and no appearance information, as input, our approach achieves superior results compared to previous state-of-the-art self-supervised methods on public benchmarks (DAVIS2016, SegTrackv2, FBMS59), while being an order of magnitude faster. On a challenging camouflage dataset (MoCA), we significantly outperform other self-supervised approaches, and are competitive with the top supervised approach, highlighting the importance of motion cues and the potential bias towards appearance in existing video segmentation models.
****

End-to-End Piece-Wise Unwarping of Document Images
Sagnik Das, Kunwar Yashraj Singh, Jon Wu, Erhan Bas, Vijay Mahadevan, Rahul Bhotika, Dimitris Samaras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4268-4277
Document unwarping attempts to undo the physical deformation of the paper and recover a 'flatbed' scanned document-image for downstream tasks such as OCR. Current state-of-the-art relies on global unwarping of the document which is not robust to local deformation changes. Moreover, a global unwarping often produces spurious warping artifacts in less warped regions to compensate for severe warps present in other parts of the document. In this paper, we propose the first end-to-end trainable piece-wise unwarping method that predicts local deformation fields and stitches them together with global information to obtain an improved unwarping. The proposed piece-wise formulation results in 4% improvement in terms of multi-scale structural similarity (MS-SSIM) and shows better performance in terms of OCR metrics, character error rate (CER) and word error rate (WER) compared to the state-of-the-art.
****

4D Cloud Scattering Tomography
Roi Ronen, Yoav Y. Schechner, Eshkol Eytan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5520-5529
We derive computed tomography (CT) of a time-varying volumetric scattering object, using a small number of moving cameras. We focus on passive tomography of dynamic clouds, as clouds have a major effect on the Earth's climate. State of the art scattering CT assumes a static object. Existing 4D CT methods rely on a linear image formation model and often on significant priors. In this paper, the angular and temporal sampling rates needed for a proper recovery are discussed. Spatiotemporal CT is achieved using gradient-based optimization, which accounts for the correlation time of the dynamic object content. We demonstrate this in physics-based simulations and on experimental real-world data.
****

Weakly Supervised Representation Learning With Coarse Labels
Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, Juhua Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10593-10601
With the development of computational power and techniques for data collection, deep learning demonstrates a superior performance over most existing algorithms on visual benchmark data sets. Many efforts have been devoted to studying the mechanism of deep learning. One important observation is that deep learning can learn the discriminative patterns from raw materials directly in a task-dependent manner. Therefore, the representations obtained by deep learning outperform hand-crafted features significantly. However, for some real-world applications, it is too expensive to collect the task-specific labels, such as visual search in online shopping. Compared to the limited availability of these task-specific labels, their coarse-class labels are much more affordable, but representations learned from them can be suboptimal for the target task. To mitigate this challenge,

we propose an algorithm to learn the fine-grained patterns for the target task, when only its coarse-class labels are available. More importantly, we provide a theoretical guarantee for this. Extensive experiments on real-world data sets demonstrate that the proposed method can significantly improve the performance of learned representations on the target task, when only coarse-class information is available for training.

****

Asymmetric Bilateral Motion Estimation for Video Frame Interpolation

Junheum Park, Chul Lee, Chang-Su Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14539-14548

We propose a novel video frame interpolation algorithm based on asymmetric bilateral motion estimation (ABME), which synthesizes an intermediate frame between two input frames. First, we predict symmetric bilateral motion fields to interpolate an anchor frame. Second, we estimate asymmetric bilateral motions fields from the anchor frame to the input frames. Third, we use the asymmetric fields to warp the input frames backward and reconstruct the intermediate frame. Last, to refine the intermediate frame, we develop a new synthesis network that generates a set of dynamic filters and a residual frame using local and global information. Experimental results show that the proposed algorithm achieves excellent performance on various datasets. The source codes and pretrained models are available at https://github.com/JunHeum/ABME.

****

LocalTrans: A Multiscale Local Transformer Network for Cross-Resolution Homography Estimation

Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, Lu Fang, Yebin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14890-14899

Cross-resolution image alignment is a key problem in multiscale gigapixel photography, which requires to estimate homography matrix using images with large resolution gap. Existing deep homography methods concatenate the input images or features, neglecting the explicit formulation of correspondences between them, which leads to degraded accuracy in cross-resolution challenges. In this paper, we consider the cross-resolution homography estimation as a multimodal problem, and propose a local transformer network embedded within a multiscale structure to explicitly learn correspondences between the multimodal inputs, namely, input images with different resolutions. The proposed local transformer adopts a local attention map specifically for each position in the feature. By combining the local transformer with the multiscale structure, the network is able to capture long-short range correspondences efficiently and accurately. Experiments on both the MS-COCO dataset and real-captured cross-resolution dataset show that the proposed network outperforms existing state-of-the-art feature-based and deep-learning-based homography estimation methods, and is able to accurately align images under 10x resolution gap.

****

De-Rendering Stylized Texts

Wataru Shimoda, Daichi Haraguchi, Seiichi Uchida, Kota Yamaguchi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1076-1085

Editing raster text is a promising but challenging task. We propose to apply text vectorization for the task of raster text editing in display media, such as posters, web pages, or advertisements. In our approach, instead of applying image transformation or generation in the raster domain, we learn a text vectorization model to parse all the rendering parameters including text, location, size, font, style, effects, and hidden background, then utilize those parameters for reconstruction and any editing task. Our text vectorization takes advantage of differentiable text rendering to accurately reproduce the input raster text in a resolution-free parametric format. We show in the experiments that our approach can successfully parse text, styling, and background information in the unified model, and produces artifact-free text editing compared to a raster baseline.

****

HRegNet: A Hierarchical Network for Large-Scale Outdoor LiDAR Point Cloud Registration
Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, Rongqi Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16014-16023
Point cloud registration is a fundamental problem in 3D computer vision. Outdoor LiDAR point clouds are typically large-scale and complexly distributed, which makes the registration challenging. In this paper, we propose an efficient hierarchical network named HRegNet for large-scale outdoor LiDAR point cloud registration. Instead of using all points in the point clouds, HRegNet performs registration on hierarchically extracted keypoints and descriptors. The overall framework combines the reliable features in deeper layer and the precise position information in shallower layers to achieve robust and precise registration. We present a correspondence network to generate correct and accurate keypoints correspondences. Moreover, bilateral consensus and neighborhood consensus are introduced for keypoints matching and novel similarity features are designed to incorporate them into the correspondence network, which significantly improves the registration performance. Besides, the whole network is also highly efficient since only a small number of keypoints are used for registration. Extensive experiments are conducted on two large-scale outdoor LiDAR point cloud datasets to demonstrate the high accuracy and efficiency of the proposed HRegNet. The project website is https://ispc-group.github.io/hregnet.
****

A Latent Transformer for Disentangled Face Editing in Images and Videos
Xu Yao, Alasdair Newson, Yann Gousseau, Pierre Hellier; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13789-13798
High quality facial image editing is a challenging problem in the movie post-production industry, requiring a high degree of control and identity preservation. Previous works that attempt to tackle this problem may suffer from the entanglement of facial attributes and the loss of the person's identity. Furthermore, many algorithms are limited to a certain task. To tackle these limitations, we propose to edit facial attributes via the latent space of a StyleGAN generator, by training a dedicated latent transformation network and incorporating explicit disentanglement and identity preservation terms in the loss function. We further introduce a pipeline to generalize our face editing to videos. Our model achieves a disentangled, controllable, and identity-preserving facial attribute editing, even in the challenging case of real (i.e., non-synthetic) images and videos. We conduct extensive experiments on image and video datasets and show that our model outperforms other state-of-the-art methods in visual quality and quantitative evaluation. Source codes are available at https://github.com/InterDigitalInc/latent-transformer.
****

AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis
Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, Juyong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5784-5794
Generating high-fidelity talking head video by fitting with the input audio sequence is a challenging problem that receives considerable attentions recently. In this paper, we address this problem with the aid of neural scene representation networks. Our method is completely different from existing methods that rely on intermediate representations like 2D landmarks or 3D face models to bridge the gap between audio input and video output. Specifically, the feature of input audio signal is directly fed into a conditional implicit function to generate a dynamic neural radiance field, from which a high-fidelity talking-head video corresponding to the audio signal is synthesized using volume rendering. Another advantage of our framework is that not only the head (with hair) region is synthesized as previous methods did, but also the upper body is generated via two individual neural radiance fields. Experimental results demonstrate that our novel framework can (1) produce high-fidelity and natural results, and (2) support free adjustment of audio signals, viewing directions, and background images. Code is ava

ilable at https://github.com/YudongGuo/AD-NeRF.
****

Graph-Based Asynchronous Event Processing for Rapid Object Recognition
Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, Guofeng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 934-943
Different from traditional video cameras, event cameras capture asynchronous events stream in which each event encodes pixel location, trigger time, and the polarity of the brightness changes. In this paper, we introduce a novel graph-based framework for event cameras, namely SlideGCN. Unlike some recent graph-based methods that use groups of events as input, our approach can efficiently process data event-by-event, unlock the low latency nature of events data while still maintaining the graph's structure internally. For fast graph construction, we develop a radius search algorithm, which better exploits the partial regular structure of event cloud against k-d tree based generic methods. Experiments show that our method reduces the computational complexity up to 100 times with respect to current graph-based methods while keeping state-of-the-art performance on object recognition. Moreover, we verify the superiority of event-wise processing with our method. When the state becomes stable, we can give a prediction with high confidence, thus making an early recognition.
****

Learning With Noisy Labels via Sparse Regularization
Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, Xiangyang Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 72-81
Learning with noisy labels is an important and challenging task for training accurate deep neural networks. However, some commonly-used loss functions, such as Cross Entropy (CE), always suffer from severe overfitting to noisy labels. Although robust loss functions have been designed, they often encounter underfitting. In this paper, we theoretically prove that any loss will be robust to noisy labels when restricting the output of a network to the set of permutations over any fixed vector. When the fixed vector is one-hot, we only need to constrain the output to be one-hot, which means a discrete image and thus zero gradients almost everywhere. This prohibits gradient-based learning of models. In this work, we introduce two sparse regularization strategies to approximate the one-hot constraint: output sharpening and l_p-norm (p\le 1). Output sharpening directly modifies the output distribution of a network to be sharp by adjusting the "temperature" parameter. l_p-norm plays the role of a regularization term to make the output to be sparse. These two simple strategies guarantee the robustness of arbitrary loss functions while not hindering the fitting ability of networks. Experiments on baseline and real-world datasets demonstrate that the sparse regularization can significantly improve the performance of commonly-used loss functions in the presence of noisy labels, and outperform state-of-the-art methods.
****

Leveraging Auxiliary Tasks With Affinity Learning for Weakly Supervised Semantic Segmentation
Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, Dan Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6984-6993
Semantic segmentation is a challenging task in the absence of densely labelled data. Only relying on class activation maps (CAM) with image-level labels provides deficient segmentation supervision. Prior works thus consider pre-trained models to produce coarse saliency maps to guide the generation of pseudo segmentation labels. However, the commonly used off-line heuristic generation process cannot fully exploit the benefits of these coarse saliency maps. Motivated by the significant inter-task correlation, we propose a novel weakly supervised multi-task framework termed as AuxSegNet, to leverage saliency detection and multi-label image classification as auxiliary tasks to improve the primary task of semantic segmentation using only image-level ground-truth labels. Inspired by their similar structured semantics, we also propose to learn a cross-task global pixel-level

affinity map from the saliency and segmentation representations. The learned cross-task affinity can be used to refine saliency predictions and propagate CAM maps to provide improved pseudo labels for both tasks. The mutual boost between pseudo label updating and cross-task affinity learning enables iterative improvements on segmentation performance. Extensive experiments demonstrate the effectiveness of the proposed auxiliary learning network structure and the cross-task affinity learning method. The proposed approach achieves state-of-the-art weakly supervised segmentation performance on the challenging PASCAL VOC 2012 and MS COCO benchmarks.

****

Improving 3D Object Detection With Channel-Wise Transformer

Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Min-Jian Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2743-2752

Though 3D object detection from point clouds has achieved rapid progress in recent years, the lack of flexible and high-performance proposal refinement remains a great hurdle for existing state-of-the-art two-stage detectors. Previous works on refining 3D proposals have relied on human-designed components such as keypoints sampling, set abstraction and multi-scale feature fusion to produce powerful 3D object representations. Such methods, however, have limited ability to capture rich contextual dependencies among points. In this paper, we leverage the high-quality region proposal network and a Channel-wise Transformer architecture to constitute our two-stage 3D object detection framework (CT3D) with minimal hand-crafted design. The proposed CT3D simultaneously performs proposal-aware embedding and channel-wise context aggregation for the point features within each proposal. Specifically, CT3D uses proposal's keypoints for spatial contextual modelling and learns attention propagation in the encoding module, mapping the proposal to point embeddings. Next, a new channel-wise decoding module enriches the query-key interaction via channel-wise re-weighting to effectively merge multi-level contexts, which contributes to more accurate object predictions. Extensive experiments demonstrate that our CT3D method has superior performance and excellent scalability. Remarkably, CT3D achieves the AP of 81.77% in the moderate car category on the KITTI test 3D detection benchmark, outperforms state-of-the-art 3D detectors.

****

CanvasVAE: Learning To Generate Vector Graphic Documents

Kota Yamaguchi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5481-5489

Vector graphic documents present visual elements in a resolution free, compact format and are often seen in creative applications. In this work, we attempt to learn a generative model of vector graphic documents. We define vector graphic documents by a multi-modal set of attributes associated to a canvas and a sequence of visual elements such as shapes, images, or texts, and train variational auto-encoders to learn the representation of the documents. We collect a new dataset of design templates from an online service that features complete document structure including occluded elements. In experiments, we show that our model, named CanvasVAE, constitutes a strong baseline for generative modeling of vector graphic documents.

****

Flow-Guided Video Inpainting With Scene Templates

Dong Lao, Peihao Zhu, Peter Wonka, Ganesh Sundaramoorthi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14599-14608

We consider the problem of filling in missing spatio-temporal regions of a video. We provide a novel flow-based solution by introducing a generative model of images in relation to the scene (without missing regions) and mappings from the scene to images. We use the model to jointly infer the scene template, a 2D representation of the scene, and the mappings. This ensures consistency of the frame-to-frame flows generated to the underlying scene, reducing geometric distortions in flow-based inpainting. The template is mapped to the missing regions in the video by a new (L2-L1) interpolation scheme, creating crisp inpaintings, reducing

common blur and distortion artifacts. We show on two benchmark datasets that our approach outperforms state-of-the-art quantitatively and in user studies.
****

Long Short View Feature Decomposition via Contrastive Video Representation Learning
Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, Mehdi Noroozi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9244-9253
Self-supervised video representation methods typically focus on the representation of temporal attributes in videos. However, the role of stationary versus non-stationary attributes is less explored: Stationary features, which remain similar throughout the video, enable the prediction of video-level action classes. Non-stationary features, which represent temporally varying attributes, are more beneficial for downstream tasks involving more fine-grained temporal understanding, such as action segmentation. We argue that a single representation to capture both types of features is sub-optimal, and propose to decompose the representation space into stationary and non-stationary features via contrastive learning from long and short views, i.e. long video sequences and their shorter sub-sequences. Stationary features are shared between the short and long views, while non-stationary features aggregate the short views to match the corresponding long view. To empirically verify our approach, we demonstrate that our stationary features work particularly well on an action recognition downstream task, while our non-stationary features perform better on action segmentation. Furthermore, we analyse the learned representations and find that stationary features capture more temporally stable, static attributes, while non-stationary features encompass more temporally varying ones.
****

TACo: Token-Aware Cascade Contrastive Learning for Video-Text Alignment
Jianwei Yang, Yonatan Bisk, Jianfeng Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11562-11572
Contrastive learning has been widely used to train transformer-based vision-language models for video-text alignment and multi-modal representation learning. This paper presents a new algorithm called Token-Aware Cascade contrastive learning (TACo) that improves contrastive learning using two novel techniques. The first is the token-aware contrastive loss which is computed by taking into account the syntactic classes of words. This is motivated by the observation that for a video-text pair, the content words in the text, such as nouns and verbs, are more likely to be aligned with the visual contents in the video than the function words. Second, a cascade sampling method is applied to generate a small set of hard negative examples for efficient loss estimation for multi-modal fusion layers. To validate the effectiveness of TACo, in our experiments we finetune pretrained models for a set of downstream tasks including text-video retrieval (YouCook2, MSR-VTT and ActivityNet), video action step localization (CrossTask), video action segmentation (COIN). Our results show that our models attain consistent improvements across different experimental settings over previous methods, setting new state-of-the-art on three public text-video retrieval benchmarks of YouCook2, MSR-VTT and ActivityNet.
****

Meta Learning on a Sequence of Imbalanced Domains With Difficulty Awareness
Zhenyi Wang, Tiehang Duan, Le Fang, Qiuling Suo, Mingchen Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8947-8957
Recognizing new objects by learning from a few labeled examples in an evolving environment is crucial to obtain excellent generalization ability for real-world machine learning systems. A typical setting across current meta learning algorithms assumes a stationary task distribution during meta training. In this paper, we explore a more practical and challenging setting where task distribution changes over time with domain shift. Particularly, we consider realistic scenarios where task distribution is highly imbalanced with domain labels unavailable in nature. We propose a kernel-based method for domain change detection and a difficulty-aware memory management mechanism that jointly considers the imbalanced doma

in size and domain importance to learn across domains continuously. Furthermore, we introduce an efficient adaptive task sampling method during meta training, which significantly reduces task gradient variance with theoretical guarantees. Finally, we propose a challenging benchmark with imbalanced domain sequences and varied domain difficulty. We have performed extensive evaluations on the proposed benchmark, demonstrating the effectiveness of our method.
****

Ranking Models in Unlabeled New Environments
Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11761-11771
Consider a scenario where we are supplied with a number of ready-to-use models trained on a certain source domain and hope to directly apply the most appropriate ones to different target domains based on the models' relative performance. Ideally we should annotate a validation set for model performance assessment on each new target environment, but such annotations are often very expensive. Under this circumstance, we introduce the problem of ranking models in unlabeled new environments. For this problem, we propose to adopt a proxy dataset that 1) is fully labeled and 2) well reflects the true model rankings in a given target environment, and use the performance rankings on the proxy sets as surrogates. We first select labeled datasets as the proxy. Specifically, datasets that are more similar to the unlabeled target domain are found to better preserve the relative performance rankings. Motivated by this, we further propose to search the proxy set by sampling images from various datasets that have similar distributions as the target. We analyze the problem and its solutions on the person re-identification (re-ID) task, for which sufficient datasets are publicly available, and show that a carefully constructed proxy set effectively captures relative performance ranking in new environments. Code is avalible at https://github.com/sxzrt/Proxy-Set.
****

Adaptive Confidence Thresholding for Monocular Depth Estimation
Hyesong Choi, Hunsang Lee, Sunkyung Kim, Sunok Kim, Seungryong Kim, Kwanghoon Sohn, Dongbo Min; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12808-12818
Self-supervised monocular depth estimation has become an appealing solution to the lack of ground truth labels, but its reconstruction loss often produces over-smoothed results across object boundaries and is incapable of handling occlusion explicitly. In this paper, we propose a new approach to leverage pseudo ground truth depth maps of stereo images generated from self-supervised stereo matching methods. The confidence map of the pseudo ground truth depth map is estimated to mitigate performance degeneration by inaccurate pseudo depth maps. To cope with the prediction error of the confidence map itself, we also leverage the threshold network that learns the threshold dynamically conditioned on the pseudo depth maps. The pseudo depth labels filtered out by the thresholded confidence map are used to supervise the monocular depth network. Furthermore, we propose the probabilistic framework that refines the monocular depth map with the help of its uncertainty map through the pixel-adaptive convolution (PAC) layer. Experimental results demonstrate superior performance to state-of-the-art monocular depth estimation methods. Lastly, we exhibit that the proposed threshold learning can also be used to improve the performance of existing confidence estimation approaches.
****

Embedding Novel Views in a Single JPEG Image
Yue Wu, Guotao Meng, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14519-14527
We propose a novel approach for embedding novel views in a single JPEG image while preserving the perceptual fidelity of the modified JPEG image and the restored novel views. We adopt the popular novel view synthesis representation of multi plane images (MPIs). Our model first encodes 32 MPI layers (totally 128 channels) into a 3-channel JPEG image that can be decoded for MPIs to render novel views

, with an embedding capacity of 1024 bits per pixel. We conducted experiments on public datasets with different novel view synthesis methods, and the results show that the proposed method can restore high-fidelity novel views from a slightly modified JPEG image. Furthermore, our method is robust to JPEG compression, color adjusting, and cropping. Our source code will be publicly available.
****

Channel Augmented Joint Learning for Visible-Infrared Recognition
Mang Ye, Weijian Ruan, Bo Du, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13567-13576
This paper introduces a powerful channel augmented joint learning strategy for the visible-infrared recognition problem. For data augmentation, most existing methods directly adopt the standard operations designed for single-modality visible images, and thus do not fully consider the imagery properties in visible to infrared matching. Our basic idea is to homogenously generate color-irrelevant images by randomly exchanging the color channels. It can be seamlessly integrated into existing augmentation operations without modifying the network, consistently improving the robustness against color variations. Incorporated with a random erasing strategy, it further greatly enriches the diversity by simulating random occlusions. For cross-modality metric learning, we design an enhanced channel-mixed learning strategy to simultaneously handle the intra- and cross-modality variations with squared difference for stronger discriminability. Besides, a channel-augmented joint learning strategy is further developed to explicitly optimize the outputs of augmented images. Extensive experiments with insightful analysis on two visible-infrared recognition tasks show that the proposed strategies consistently improve the accuracy. Without auxiliary information, it improves the state-of-the-art Rank-1/mAP by 14.59%/13.00% on the large-scale SYSU-MM01 dataset.
****

OMNet: Learning Overlapping Mask for Partial-to-Partial Point Cloud Registration
Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, Bing Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3132-3141
Point cloud registration is a key task in many computational fields. Previous correspondence matching based methods require the inputs to have distinctive geometric structures to fit a 3D rigid transformation according to point-wise sparse feature matches. However, the accuracy of transformation heavily relies on the quality of extracted features, which are prone to errors with respect to partiality and noise. In addition, they can not utilize the geometric knowledge of all the overlapping regions. On the other hand, previous global feature based approaches can utilize the entire point cloud for the registration, however they ignore the negative effect of non-overlapping points when aggregating global features. In this paper, we present OMNet, a global feature based iterative network for partial-to-partial point cloud registration. We learn overlapping masks to reject non-overlapping regions, which converts the partial-to-partial registration to the registration of the same shape. Moreover, the previously used data is sampled only once from the CAD models for each object, resulting in the same point clouds for the source and reference. We propose a more practical manner of data generation where a CAD model is sampled twice for the source and reference, avoiding the previously prevalent over-fitting issue. Experimental results show that our method achieves state-of-the-art performance compared to traditional and deep learning based methods. Code is available at https://github.com/megvii-research/OMNet.
****

Learning Skeletal Graph Neural Networks for Hard 3D Pose Estimation
Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, Qiang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11436-11445
Various deep learning techniques have been proposed to solve the single-view 2D-to-3D pose estimation problem. While the average prediction accuracy has been improved significantly over the years, the performance on hard poses with depth ambiguity, self-occlusion, and complex or rare poses is still far from satisfactor

y. In this work, we target these hard poses and present a novel skeletal GNN lea rning solution. To be specific, we propose a hop-aware hierarchical channel-sque ezing fusion layer to effectively extract relevant information from neighboring nodes while suppressing undesired noises in GNN learning. In addition, we propos e a temporal-aware dynamic graph construction procedure that is robust and effec tive for 3D pose estimation. Experimental results on the Human3.6M dataset show that our solution achieves a 10.3% average prediction accuracy improvement and g reatly improves on hard poses over state-of-the-art techniques. We further apply the proposed technique on the skeleton-based action recognition task and also a chieve state-of-the-art performance.

****

Recursively Conditional Gaussian for Ordinal Unsupervised Domain Adaptation
Xiaofeng Liu, Site Li, Yubin Ge, Pengyi Ye, Jane You, Jun Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 764-773
The unsupervised domain adaptation (UDA) has been widely adopted to alleviate th e data scalability issue, while the existing works usually focus on classifying independently discrete labels. However, in many tasks (e.g., medical diagnosis), the labels are discrete and successively distributed. The UDA for ordinal class ification requires inducing non-trivial ordinal distribution prior to the latent space. Target for this, the partially ordered set (poset) is defined for constr aining the latent vector. Instead of the typically i.i.d. Gaussian latent prior, in this work, a recursively conditional Gaussian (RCG) set is adapted for order ed constraint modeling, which admits a tractable joint distribution prior. Furth ermore, we are able to control the density of content vector that violates the p oset constraints by a simple "three-sigma rule". We explicitly disentangle the c ross-domain images into a shared ordinal prior induced ordinal content space and two separate source/target ordinal-unrelated spaces, and the self-training is w orked on the shared space exclusively for ordinal-aware domain alignment. Extens ive experiments on UDA medical diagnoses and facial age estimation demonstrate i ts effectiveness.

****

Learning Anchored Unsigned Distance Functions With Gradient Direction Alignment for Single-View Garment Reconstruction
Fang Zhao, Wenhao Wang, Shengcai Liao, Ling Shao; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2021, pp. 12674-12683
While single-view 3D reconstruction has made significant progress benefiting fro m deep shape representations in recent years, garment reconstruction is still no t solved well due to open surfaces, diverse topologies and complex geometric det ails. In this paper, we propose a novel learnable Anchored Unsigned Distance Fun ction (AnchorUDF) representation for 3D garment reconstruction from a single ima ge. AnchorUDF represents 3D shapes by predicting unsigned distance fields (UDFs) to enable open garment surface modeling at arbitrary resolution. To capture div erse garment topologies, AnchorUDF not only computes pixel-aligned local image f eatures of query points, but also leverages a set of anchor points located aroun d the surface to enrich 3D position features for query points, which provides st ronger 3D space context for the distance function. Furthermore, in order to obta in more accurate point projection direction at inference, we explicitly align th e spatial gradient direction of AnchorUDF with the ground-truth direction to the surface during training. Extensive experiments on two public 3D garment dataset s, i.e., MGN and Deep Fashion3D, demonstrate that AnchorUDF achieves the state-o f-the-art performance on single-view garment reconstruction. Code is available a t https://github.com/zhaofang0627/AnchorUDF.

****

TeachText: CrossModal Generalized Distillation for Text-Video Retrieval
Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisser man, Samuel Albanie, Yang Liu; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 11583-11593
In recent years, considerable progress on the task of text-video retrieval has b een achieved by leveraging large-scale pretraining on visual and audio datasets to construct powerful video encoders. By contrast, despite the natural symmetry,

the design of effective algorithms for exploiting large-scale language pretrain ing remains under-explored. In this work, we are the first to investigate the de sign of such algorithms and propose a novel generalized distillation method,Teac hText, which leverages complementary cues from multiple text encoders to provide an enhanced supervisory signal to the retrieval model. Moreover, we extend our method to video side modalities and show that we can effectively reduce the numb er of used modalities at test time without compromising performance. Our approac h advances the state of the art on several video retrieval benchmarks by a signi ficant margin and adds no computational overhead at test time. Last but not leas t, we show an effective application of our method for eliminating noise from ret rieval datasets. Code and data can be found at https://www.robots.ox.ac.uk/ vgg/ research/teachtext/.

****

Geometry Uncertainty Projection Network for Monocular 3D Object Detection

Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, Wanl i Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2021, pp. 3111-3121

Monocular 3D object detection has received increasing attention due to the wide application in autonomous driving. Existing works mainly focus on introducing ge ometry projection to predict depth priors for each object. Despite their impress ive progress, these methods neglect the geometry leverage effect of the projecti on process, which leads to uncontrollable inferences and damage the training eff iciency. In this paper, we propose a Geometry Uncertainty Projection Network (GU P Net) to handle these problems, which can guide the model to learn more reliabl e depth outputs. The overall framework combines the uncertainty inference and th e hierarchical task learning to reduce the negative effects of the geometry leve rage. Specifically, an Uncertainty Geometry Projection module is proposed to obt ain the geometry guided uncertainty of the inferred depth, which can not only be nefit the geometry learning but also provide more reliable depth inferences to r educe the uncontrollableness caused by the geometry leverage. Besides, to reduce the instability in the training process caused by the geometry leverage effect, we propose a Hierarchical Task Learning strategy to control the overall optimiz ation process. This learning algorithm can monitor the situation of each task th rough a well designed learning situation indicator and adaptively assign the pro per loss weights for different tasks according to their learning situation and t he hierarchical structure, which can significantly improve the stability and the efficiency of the training process. Extensive experiments demonstrate the effec tiveness of the proposed method.The overall model can infer more reliable depth and location information than existing methods, which achieves the state-of-the-art performance on the KITTI benchmark.

****

OVANet: One-vs-All Network for Universal Domain Adaptation

Kuniaki Saito, Kate Saenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9000-9009

Universal Domain Adaptation (UNDA) aims to handle both domain-shift and category -shift between two datasets, where the main challenge is to transfer knowledge w hile rejecting "unknown" classes which are absent in the labeled source data but present in the unlabeled target data. Existing methods manually set a threshold to reject "unknown" samples based on validation or a pre-defined ratio of "unkn own" samples, but this strategy is not practical. In this paper, we propose a me thod to learn the threshold using source samples and to adapt it to the target d omain. Our idea is that a minimum inter-class distance in the source domain shou ld be a good threshold to decide between "known" or "unknown" in the target. To learn the inter- and intra-class distance, we propose to train a one-vs-all clas sifier for each class using labeled source data. Then, we adapt the open-set cla ssifier to the target domain by minimizing class entropy. The resulting framewor k is the simplest of all baselines of UNDA and is insensitive to the value of a hyper-parameter, yet outperforms baselines with a large margin.

****

A Hybrid Frequency-Spatial Domain Model for Sparse Image Reconstruction in Scann

ing Transmission Electron Microscopy
Bintao He, Fa Zhang, Huanshui Zhang, Renmin Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2682-2691
Scanning transmission electron microscopy (STEM) is a powerful technique in high-resolution atomic imaging of materials. Decreasing scanning time and reducing electron beam exposure with an acceptable signal-to-noise results are two popular research aspects when applying STEM to beam-sensitive materials. Specifically, partially sampling with fixed electron doses is one of the most important solutions, and then the lost information is restored by computational methods. Following successful applications of deep learning in image in-painting, we have developed an encoder-decoder network to reconstruct STEM images in extremely sparse sampling case. In our model, we combine both local pixel information from convolution operators and global texture features, by applying specific filter operations on frequency domain to acquire initial reconstruction and global structure prior. Our method can effectively restore texture structures and be robust in different sampling ratios with Poisson noise. A comprehensive study demonstrates that our method gains about 50% performance enhancement in comparison with the state-of-art methods. Code is available at https://github.com/icthrm/Sparse-Sampling-Reconstruction.
****
Attentional Pyramid Pooling of Salient Visual Residuals for Place Recognition
Guohao Peng, Jun Zhang, Heshan Li, Danwei Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 885-894
The core of visual place recognition (VPR) lies in how to identify task-relevant visual cues and embed them into discriminative representations. Focusing on these two points, we propose a novel encoding strategy named Attentional Pyramid Pooling of Salient Visual Residuals (APPSVR). It incorporates three types of attention modules to model the saliency of local features in individual, spatial and cluster dimensions respectively. (1) To inhibit task-irrelevant local features, a semantic-reinforced local weighting scheme is employed for local feature refinement; (2) To leverage the spatial context, an attentional pyramid structure is constructed to adaptively encode regional features according to their relative spatial saliency; (3) To distinguish the different importance of visual clusters to the task, a parametric normalization is proposed to adjust their contribution to image descriptor generation. Experiments demonstrate APPSVR outperforms the existing techniques and achieves a new state-of-the-art performance on VPR benchmark datasets. The visualization shows the saliency map learned in a weakly supervised manner is largely consistent with human cognition.
****
Learning With Noisy Labels for Robust Point Cloud Segmentation
Shuquan Ye, Dongdong Chen, Songfang Han, Jing Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6443-6452
Point cloud segmentation is a fundamental task in 3D. Despite recent progress on point cloud segmentation with the power of deep networks, current deep learning methods based on the clean label assumptions may fail with noisy labels. Yet, object class labels are often mislabeled in real-world point cloud datasets. In this work, we take the lead in solving this issue by proposing a novel Point Noise-Adaptive Learning (PNAL) framework. Compared to existing noise-robust methods on image tasks, our PNAL is noise-rate blind, to cope with the spatially variant noise rate problem specific to point clouds. Specifically, we propose a novel point-wise confidence selection to obtain reliable labels based on the historical predictions of each point. A novel cluster-wise label correction is proposed with a voting strategy to generate the best possible label taking the neighbor point correlations into consideration. We conduct extensive experiments to demonstrate the effectiveness of PNAL on both synthetic and real-world noisy datasets. In particular, even with 60% symmetric noisy labels, our proposed method produces much better results than its baseline counterpart without PNAL and is comparable to the ideal upper bound trained on a completely clean dataset. Moreover, we fully re-labeled the validation set of a popular but noisy real-world scene dataset ScanNetV2 to make it clean, for rigorous experiment and future research. Our

code and data will be released.
****
Where Are You Heading? Dynamic Trajectory Prediction With Expert Goal Examples
He Zhao, Richard P. Wildes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7629-7638
Goal-conditioned approaches recently have been found very useful to human trajectory prediction, when adequate goal estimates are provided. Yet, goal inference is difficult in itself and often incurs extra learning efforts. We propose to predict pedestrian trajectories via the guidance of goal expertise, which can be obtained with modest expense through a novel goal-search mechanism on already seen training examples. There are three key contributions in our study. First, we devise a framework that exploits the nearest examples for high-quality goal position inquiry. This approach naturally considers multi-modality, physical constraints, compatibility with existing methods and is model-free; it therefore does not require additional learning efforts typical in goal inference. Second, we present an end-to-end trajectory predictor that can efficiently associate goal retrievals to past motion information and dynamically infer possible future trajectories. Third, with these two novel techniques in hand, we conduct a series of experiments on two broadly explored datasets (SDD and ETH/UCY) and show that our approach surpasses previous state-of-the-art performance by notable margins and reduces the need for additional parameters.
****
Planar Surface Reconstruction From Sparse Views
Linyi Jin, Shengyi Qian, Andrew Owens, David F. Fouhey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12991-13000
The paper studies planar surface reconstruction of indoor scenes from two views with unknown camera poses. While prior approaches have successfully created object-centric reconstructions of many scenes, they fail to exploit other structures, such as planes, which are typically the dominant components of indoor scenes. In this paper, we reconstruct planar surfaces from multiple views, while jointly estimating camera pose. Our experiments demonstrate that our method is able to advance the state of the art of reconstruction from sparse views, on challenging scenes from Matterport3D.
****
Dynamic Divide-and-Conquer Adversarial Training for Robust Semantic Segmentation
Xiaogang Xu, Hengshuang Zhao, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7486-7495
Adversarial training is promising for improving robustness of deep neural networks towards adversarial perturbations, especially on the classification task. The effect of this type of training on semantic segmentation, contrarily, just commences. We make the initial attempt to explore the defense strategy on semantic segmentation by formulating a general adversarial training procedure that can perform decently on both adversarial and clean samples. We propose a dynamic divide-and-conquer adversarial training (DDC-AT) strategy to enhance the defense effect, by setting additional branches in the target model during training, and dealing with pixels with diverse properties towards adversarial perturbation. Our dynamical division mechanism divides pixels into multiple branches automatically. Note all these additional branches can be abandoned during inference and thus leave no extra parameter and computation cost. Extensive experiments with various segmentation models are conducted on PASCAL VOC 2012 and Cityscapes datasets, in which DDC-AT yields satisfying performance under both white- and black-box attack. The code is available at https://github.com/dvlab-research/Robust-Semantic-Segmentation.
****
MixMix: All You Need for Data-Free Compression Are Feature and Data Mixing
Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, Shi Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4410-4419
User data confidentiality protection is becoming a rising challenge in the present deep learning research. Without access to data, conventional data-driven mode

l compression faces a higher risk of performance degradation. Recently, some wor ks propose to generate images from a specific pretrained model to serve as train ing data. However, the inversion process only utilizes biased feature statistics stored in one model and is from low-dimension to high-dimension. As a consequen ce, it inevitably encounters the difficulties of generalizability and inexact in version, which leads to unsatisfactory performance. To address these problems, w e propose MixMix based on two simple yet effective techniques: (1) Feature Mixin g: utilizes various models to construct a universal feature space for generalize d inversion; (2) Data Mixing: mixes the synthesized images and labels to generat e exact label information. We prove the effectiveness of MixMix from both theore tical and empirical perspectives. Extensive experiments show that MixMix outperf orms existing methods on the mainstream compression tasks, including quantizatio n, knowledge distillation and pruning. Specifically, MixMix achieves up to 4% an d 20% accuracy uplift on quantization and pruning, respectively, compared to exi sting data-free compression work.

****

VidTr: Video Transformer Without Convolutions

Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Che n, Ivan Marsic, Joseph Tighe; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2021, pp. 13577-13587

We introduce Video Transformer (VidTr) with separable-attention for video classi fication. Comparing with commonly used 3D networks, VidTr is able to aggregate s patio-temporal information via stacked attentions and provide better performance with higher efficiency. We first introduce the vanilla video transformer and sh ow that the transformer module is able to perform spatio-temporal modeling from raw pixels, but with heavy memory usage. We then present VidTr which reduces the memory cost by 3.3xwhile keeping the same performance. To further optimize the model, we propose the standard deviation based topK pooling for attention, which reduces the computation by dropping non-informative features along temporal dim ension. VidTr achieves state-of-the-art performance on five commonly used datase ts with lower computational requirements, showing both the efficiency and effect iveness of our design. Finally, error analysis and visualization show that VidTr is especially good at predicting actions that require long-term temporal reason ing.

****

LocTex: Learning Data-Efficient Visual Representations From Localized Textual Su pervision

Zhijian Liu, Simon Stent, Jie Li, John Gideon, Song Han; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 2167-2176

Computer vision tasks such as object detection and semantic/instance segmentatio n rely on the painstaking annotation of large training datasets. In this paper, we propose LocTex that takes advantage of the low-cost localized textual annotat ions (i.e., captions and synchronized mouse-over gestures) to reduce the annotat ion effort. We introduce a contrastive pre-training framework between images and captions and propose to supervise the cross-modal attention map with rendered m ouse traces to provide coarse localization signals. Our learned visual features capture rich semantics (from free-form captions) and accurate localization (from mouse traces), which are very effective when transferred to various downstream vision tasks. Compared with ImageNet supervised pre-training, LocTex can reduce the size of the pre-training dataset by 10x or the target dataset by 2x while ac hieving comparable or even improved performance on COCO instance segmentation. W hen provided with the same amount of annotations, LocTex achieves around 4% high er accuracy than the previous state-of-the-art "vision+language" pre-training ap proach on the task of PASCAL VOC image classification.

****

Weakly Supervised Segmentation of Small Buildings With Point Labels

Jae-Hun Lee, ChanYoung Kim, Sanghoon Sull; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 7406-7415

Most supervised image segmentation methods require delicate and time-consuming p ixel-level labeling of building or objects, especially for small objects. In thi

s paper, we present a weakly supervised segmentation network for aerial/satellite images, separately considering small and large objects. First, we propose a simple point labeling method for small objects, while large objects are fully labeled. Then, we present a segmentation network trained with a small object mask to separate small and large objects in the loss function. During training, we employ a memory bank to cope with the limited number of point labels. Experiments results with three public datasets demonstrate the feasibility of our approach.
****

Online Knowledge Distillation for Efficient Pose Estimation
Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, Zhigeng Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11740-11750

Existing state-of-the-art human pose estimation methods require heavy computational resources for accurate predictions. One promising technique to obtain an accurate yet lightweight pose estimator is knowledge distillation, which distills the pose knowledge from a powerful teacher model to a less-parameterized student model. However, existing pose distillation works rely on a heavy pre-trained estimator to perform knowledge transfer and require a complex two-stage learning procedure. In this work, we investigate a novel Online Knowledge Distillation framework by distilling Human Pose structure knowledge in a one-stage manner to guarantee the distillation efficiency, termed OKDHP. Specifically, OKDHP trains a single multi-branch network and acquires the predicted heatmaps from each, which are then assembled by a Feature Aggregation Unit (FAU) as the target heatmaps to teach each branch in reverse. Instead of simply averaging the heatmaps, FAU which consists of multiple parallel transformations with different receptive fields, leverages the multi-scale information, thus obtains target heatmaps with higher-quality. Specifically, the pixel-wise Kullback-Leibler (KL) divergence is utilized to minimize the discrepancy between the target heatmaps and the predicted ones, which enables the student network to learn the implicit keypoint relationship. Besides, an unbalanced OKDHP scheme is introduced to customize the student networks with different compression rates. The effectiveness of our approach is demonstrated by extensive experiments on two common benchmark datasets, MPII and COCO.
****

HAA500: Human-Centric Atomic Action Dataset With Curated Videos
Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13465-13474

We contribute HAA500, a manually annotated human-centric atomic action dataset for action recognition on 500 classes with over 591K labeled frames. To minimize ambiguities in action classification, HAA500 consists of highly diversified classes of fine-grained atomic actions, where only consistent actions fall under the same label, e.g., ""Baseball Pitching"" vs ""Free Throw in Basketball"". Thus HAA500 is different from existing atomic action datasets, where coarse-grained atomic actions were labeled with coarse action-verbs such as ""Throw"". HAA500 has been carefully curated to capture the precise movement of human figures with little class-irrelevant motions or spatio-temporal label noises. The advantages of HAA500 are fourfold: 1) human-centric actions with a high average of 69.7% detectable joints for the relevant human poses; 2) high scalability since adding a new class can be done under 20-60 minutes; 3) curated videos capturing essential elements of an atomic action without irrelevant frames; 4) fine-grained atomic action classes. Our extensive experiments including cross-data validation using datasets collected in the wild demonstrate the clear benefits of human-centric and atomic characteristics of HAA500, which enable training even a baseline deep learning model to improve prediction by attending to atomic human poses. We detail the HAA500 dataset statistics and collection methodology and compare quantitatively with existing action recognition datasets.
****

Efficient Large Scale Inlier Voting for Geometric Vision Problems
Dror Aiger, Simon Lynen, Jan Hosang, Bernhard Zeisl; Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), 2021, pp. 3243-3251
Outlier rejection and equivalently inlier set optimization is a key ingredient in numerous applications in computer vision such as filtering point-matches in camera pose estimation or plane and normal estimation in point clouds. Several approaches exist, yet at large scale we face a combinatorial explosion of possible solutions and state-of-the-art methods like RANSAC, Hough transform or Branch&Bound require a minimum inlier ratio or prior knowledge to remain practical. In fact, for problems such as camera posing in very large scenes these approaches become useless as they have exponential runtime growth if these conditions aren't met. To approach the problem we present a efficient and general algorithm for outlier rejection based on "intersecting" k-dimensional surfaces in Rd . We provide a recipe for casting a variety of geometric problems as finding a point in Rd which maximizes the number of nearby surfaces (and thus inliers). The resulting algorithm has linear worst-case complexity with a better runtime dependency in the approximation factor than competing algorithms while not requiring domain specific bounds. This is achieved by introducing a space decomposition scheme that bounds the number of computations by successively rounding and grouping samples. Our recipe (and open-source code) enables anybody to derive such fast approaches to new problems across a wide range of domains. We demonstrate the versatility of the approach on several camera posing problems with a high number of matches at low inlier ratio achieving state-of-the-art results at significantly lower processing times.
****

From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting
Karttikeya Mangalam, Yang An, Harshayu Girase, Jitendra Malik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15233-15242
Human trajectory forecasting is an inherently multimodal problem. Uncertainty in future trajectories stems from two sources: (a) sources that are known to the agent but unknown to the model, such as long term goals and (b) sources that are unknown to both the agent & the model, such as intent of other agents & irreducible randomness in decisions. We propose to factorize this uncertainty into its epistemic & aleatoric sources. We model the epistemic uncertainty through multimodality in long term goals and the aleatoric uncertainty through multimodality in waypoints & paths. To exemplify this dichotomy, we also propose a novel long term trajectory forecasting setting, with prediction horizons upto a minute, upto an order of magnitude longer than prior works. Finally, we present Y-net, a scene compliant trajectory forecasting network that exploits the proposed epistemic & aleatoric structure for diverse trajectory predictions across long prediction horizons. Y-net significantly improves previous state-of-the-art performance on both (a) The short prediction horizon setting on the Stanford Drone (31.7% in FDE) & ETH/UCY datasets (7.4% in FDE) and (b) The proposed long horizon setting on the re-purposed Stanford Drone & Intersection Drone datasets. Code is available at: https://karttikeya.github.io/publication/ynet/
****

DiscoBox: Weakly Supervised Instance Segmentation and Semantic Correspondence From Box Supervision
Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S. Davis, Anima Anandkumar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3406-3416
We introduce DiscoBox, a novel framework that jointly learns instance segmentation and semantic correspondence using bounding box supervision. Specifically, we propose a self-ensembling framework where instance segmentation and semantic correspondence are jointly guided by a structured teacher in addition to the bounding box supervision. The teacher is a structured energy model incorporating a pairwise potential and a cross-image potential to model the pairwise pixel relationships both within and across the boxes. Minimizing the teacher energy simultaneously yields refined object masks and dense correspondences between intra-class objects, which are taken as pseudo-labels to supervise the task network and provide positive/negative correspondence pairs for dense contrastive learning. We sho

w a symbiotic relationship where the two tasks mutually benefit from each other. Our best model achieves 37.9% AP on COCO instance segmentation, surpassing prio r weakly supervised methods and is competitive to supervised methods. We also ob tain state of the art weakly supervised results on PASCAL VOC12 and PF-PASCAL wi th real-time inference.

****

VENet: Voting Enhancement Network for 3D Object Detection

Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Dening Lu, Mingqiang Wei, Jun Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3712-3721

Hough voting, as has been demonstrated in VoteNet, is effective for 3D object de tection, where voting is a key step. In this paper, we propose a novel VoteNet-b ased 3D detector with vote enhancement to improve the detection accuracy in clut tered indoor scenes. It addresses the limitations of current voting schemes, i.e ., votes from neighboring objects and background have significant negative impac ts Specifically, before voting, we replace the classic MLP with the proposed Att entive MLP (AMLP) in the backbone network to get better feature description of s eed points. During voting, we design a new vote attraction loss (VALoss) to enfo rce vote centers to locate closely and compactly to the corresponding object cen ters. After voting, we then devise a vote weighting module to integrate the fore ground/background prediction into the vote aggregation process to enhance the ca pability of the original VoteNet to handle noise from background voting. The thr ee proposed strategies all contribute to more effective voting and improved perf ormance, resulting in a novel 3D object detector, termed VENet. Experiments show that our method outperforms state-of-the-art methods on benchmark datasets. Abl ation studies demonstrate the effectiveness of the proposed components.

****

Intrinsic-Extrinsic Preserved GANs for Unsupervised 3D Pose Transfer

Haoyu Chen, Hao Tang, Henglin Shi, Wei Peng, Nicu Sebe, Guoying Zhao; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8630-8639

With the strength of deep generative models, 3D pose transfer regains intensive research interests in recent years. Existing methods mainly rely on a variety of constraints to achieve the pose transfer over 3D meshes, e.g., the need for man ually encoding for shape and pose disentanglement. In this paper, we present an unsupervised approach to conduct the pose transfer between any arbitrate given 3 D meshes. Specifically, a novel Intrinsic-Extrinsic Preserved Generative Adversa rial Network (IEP-GAN) is presented for both intrinsic (i.e., shape) and extrins ic (i.e., pose) information preservation. Extrinsically, we propose a co-occurre nce discriminator to capture the structural/pose invariance from distinct Laplac ians of the mesh. Meanwhile, intrinsically, a local intrinsic-preserved loss is introduced to preserve the geodesic priors while avoiding heavy computations. At last, we show the possibility of using IEP-GAN to manipulate 3D human meshes in various ways, including pose transfer, identity swapping and pose interpolation with latent code vector arithmetic. The extensive experiments on various 3D dat asets of humans, animals and hands qualitatively and quantitatively demonstrate the generality of our approach. Our proposed model produces better results and i s substantially more efficient compared to recent state-of-the-art methods. Code is available: https://github.com/mikecheninoulu/Unsupervised_IEPGAN

****

Aggregation With Feature Detection

Shuyang Sun, Xiaoyu Yue, Xiaojuan Qi, Wanli Ouyang, Victor Adrian Prisacariu, Ph ilip H.S. Torr; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 527-536

Aggregating features from different depths of a network is widely adopted to imp rove the network capability. Lots of modern architectures are equipped with ski p connections, which actually makes the feature aggregation happen in all these networks. Since different features tell different semantic meanings, there are i nconsistencies and incompatibilities to be solved. However, existing works naive ly blend deep features via element-wise summation or concatenation with a convol

ution behind. Better feature aggregation method beyond summation or concatenatio n is rarely explored. In this paper, given two layers of features to be aggregat ed together, we first detect and identify where and what needs to be updated in one layer, then replace the feature at the identified location with the informat ion of the other layer. This process, which we call DEtect-rePLAce (DEPLA), enab les us to avoid inconsistent patterns while keeping useful information in the me rged outputs. Experimental results demonstrate our method largely boosts multipl e baselines e.g. ResNet, FishNet and FPN on three major vision tasks including I mageNet classification, MS COCO object detection and instance segmentation.
****

Multi-Echo LiDAR for 3D Object Detection
Yunze Man, Xinshuo Weng, Prasanna Kumar Sivakumar, Matthew O'Toole, Kris M. Kita ni; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 3763-3772
LiDAR sensors can be used to obtain a wide range of measurement signals other th an a simple 3D point cloud, and those signals can be leveraged to improve percep tion tasks like 3D object detection. A single laser pulse can be partially refle cted by multiple objects along its path, resulting in multiple measurements call ed echoes. Multi-echo measurement can provide information about object contours and semi-transparent surfaces which can be used to better identify and locate ob jects. LiDAR can also measure surface reflectance (intensity of laser pulse retu rn), as well as ambient light of the scene (sunlight reflected by objects). Thes e signals are already available in commercial LiDAR devices but have not been us ed in most LiDAR-based detection models. We present a 3D object detection model which leverages the full spectrum of measurement signals provided by LiDAR. Firs t, we propose a multi-signal fusion (MSF) module to combine (1) the reflectance and ambient features extracted with a 2D CNN, and (2) point cloud features extra cted using a 3D graph neural network (GNN). Second, we propose a multi-echo aggr egation (MEA) module to combine the information encoded in different set of echo points. Compared with traditional single echo point cloud methods, our proposed multi-signal LiDAR Detector (MSLiD) extracts richer context information from a wider range of sensing measurements and achieves more accurate 3D object detecti on. Experiments show that by incorporating the multi-modality of LiDAR, our meth od outperforms the state-of-the-art by up to relatively 9.1%.
****

Self-Regulation for Semantic Segmentation
Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, Qianru Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 69 53-6963
In this paper, we seek reasons for the two major failure cases in Semantic Segme ntation (SS): 1) missing small objects or minor object parts, and 2) mislabeling minor parts of large objects as wrong classes. We have an interesting finding t hat Failure-1 is due to the underuse of detailed features and Failure-2 is due t o the underuse of visual contexts. To help the model learn a better trade-off, w e introduce several Self-Regulation (SR) losses for training SS neural networks. By "self", we mean that the losses are from the model per se without using any additional data or supervision. By applying the SR losses, the deep layer featur es are regulated by the shallow ones to preserve more details; meanwhile, shallo w layer classification logits are regulated by the deep ones to capture more sem antics. We conduct extensive experiments on both weakly and fully supervised SS tasks, and the results show that our approach consistently surpasses the baselin es. We also validate that SR losses are easy to implement in various state-of-th e-art SS models, e.g., SPGNet and OCRNet, incurring little computational overhea d during training and none for testing
****

Skeleton2Mesh: Kinematics Prior Injected Unsupervised Human Mesh Recovery
Zhenbo Yu, Junjie Wang, Jingwei Xu, Bingbing Ni, Chenglong Zhao, Minsi Wang, Wen jun Zhang; Proceedings of the IEEE/CVF International Conference on Computer Visi on (ICCV), 2021, pp. 8619-8629
In this paper, we decouple unsupervised human mesh recovery into the well-studie

d problems of unsupervised 3D pose estimation, and human mesh recovery from esti mated 3D skeletons, focusing on the latter task. The challenges of the latter ta sk are two folds: (1) pose failure (i.e., pose mismatching -- different skeleton definitions in dataset and SMPL , and pose ambiguity -- endpoints have arbitrar y joint angle configurations for the same 3D joint coordinates). (2) shape ambig uity (i.e., the lack of shape constraints on body configuration). To address the se issues, we propose Skeleton2Mesh, a novel lightweight framework that recovers human mesh from a single image. Our Skeleton2Mesh contains three modules, i.e., Differentiable Inverse Kinematics (DIK), Pose Refinement (PR) and Shape Refinem ent (SR) modules. DIK is designed to transfer 3D rotation from estimated 3D skel etons, which relies on a minimal set of kinematics prior knowledge. Then PR and SR modules are utilized to tackle the pose ambiguity and shape ambiguity respect ively. All three modules can be incorporated into Skeleton2Mesh seamlessly via a n end-to-end manner. Furthermore, we utilize an adaptive joint regressor to alle viate the effects of skeletal topology from different datasets. Results on the H uman3.6M dataset for human mesh recovery demonstrate that our method improves up on the previous unsupervised methods by 32.6% under the same setting. Qualitativ e results on in-the-wild datasets exhibit that the recovered 3D meshes are natur al, realistic. Our project is available at https://sites.google.com/view/skeleto n2mesh.

****

ReCU: Reviving the Dead Weights in Binary Neural Networks

Zihan Xu, Mingbao Lin, Jianzhuang Liu, Jie Chen, Ling Shao, Yue Gao, Yonghong Ti an, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2021, pp. 5198-5208

Binary neural networks (BNNs) have received increasing attention due to their su perior reductions of computation and memory. Most existing works focus on either lessening the quantization error by minimizing the gap between the full-precisi on weights and their binarization or designing a gradient approximation to mitig ate the gradient mismatch, while leaving the "dead weights" untouched. This lead s to slow convergence when training BNNs. In this paper, for the first time, we explore the influence of "dead weights" which refer to a group of weights that a re barely updated during the training of BNNs, and then introduce rectified clam p unit (ReCU) to revive the "dead weights" for updating. We prove that reviving the "dead weights" by ReCU can result in a smaller quantization error. Besides, we also take into account the information entropy of the weights, and then mathe matically analyze why the weight standardization can benefit BNNs. We demonstrat e the inherent contradiction between minimizing the quantization error and maxim izing the information entropy, and then propose an adaptive exponential schedule r to identify the range of the "dead weights". By considering the "dead weights" , our method offers not only faster BNN training, but also state-of-the-art perf ormance on CIFAR-10 and ImageNet, compared with recent methods. Code can be avai lable at https://github.com/z-hXu/ReCU.

****

Membership Inference Attacks Are Easier on Difficult Problems

Avital Shafran, Shmuel Peleg, Yedid Hoshen; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2021, pp. 14820-14829

Membership inference attacks (MIA) try to detect if data samples were used to tr ain a neural network model, e.g. to detect copyright abuses. We show that models with higher dimensional input and output are more vulnerable to MIA, and addres s in more detail models for image translation and semantic segmentation, includi ng medical image segmentation. We show that reconstruction-errors can lead to ve ry effective MIA attacks as they are indicative of memorization. Unfortunately, reconstruction error alone is less effective at discriminating between non-predi ctable images used in training and easy to predict images that were never seen b efore. To overcome this, we propose using a novel predictability error that can be computed for each sample, and its computation does not require a training set . Our membership error, obtained by subtracting the predictability error from th e reconstruction error, is shown to achieve high MIA accuracy on an extensive nu mber of benchmarks.

****
Auxiliary Tasks and Exploration Enable ObjectGoal Navigation
Joel Ye, Dhruv Batra, Abhishek Das, Erik Wijmans; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16117-16126
ObjectGoal Navigation (ObjectNav) is an embodied task wherein agents are to navigate to an object instance in an unseen environment. Prior works have shown that end-to-end ObjectNav agents that use vanilla visual and recurrent modules, e.g. a CNN+RNN, perform poorly due to overfitting and sample inefficiency. This has motivated current state-of-the-art methods to mix analytic and learned components and operate on explicit spatial maps of the environment. We instead re-enable a generic learned agent by adding auxiliary learning tasks and an exploration reward. Our agents achieve 24.5% success and 8.1% SPL, a 37% and 8% relative improvement over prior state-of-the-art, respectively, on the Habitat ObjectNav Challenge. From our analysis, we propose that agents will act to simplify their visual inputs so as to smooth their RNN dynamics, and that auxiliary tasks reduce overfitting by minimizing effective RNN dimensionality; i.e. a performant ObjectNav agent that must maintain coherent plans over long horizons does so by learning smooth, low-dimensional recurrent dynamics.
****
Semantic-Embedded Unsupervised Spectral Reconstruction From Single RGB Images in the Wild
Zhiyu Zhu, Hui Liu, Junhui Hou, Huanqiang Zeng, Qingfu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2279-2288
This paper investigates the problem of reconstructing hyperspectral (HS) images from single RGB images captured by commercial cameras, without using paired HS and RGB images during training. To tackle this challenge, we propose a new lightweight and end-to-end learning-based framework. Specifically, on the basis of the intrinsic imaging degradation model of RGB images from HS images, we progressively spread the differences between input RGB images and re-projected RGB images from recovered HS images via effective unsupervised camera spectral response function estimation. To enable the learning without paired ground-truth HS images as supervision, we adopt the adversarial learning manner and boost it with a simple yet effective L1 gradient clipping scheme. Besides, we embed the semantic information of input RGB images to locally regularize the unsupervised learning, which is expected to promote pixels with identical semantics to have consistent spectral signatures. In addition to conducting quantitative experiments over two widely-used datasets for HS image reconstruction from synthetic RGB images, we also evaluate our method by applying recovered HS images from real RGB images to HS-based visual tracking. Extensive results show that our method significantly outperforms state-of-the-art unsupervised methods and even exceeds the latest supervised method under some settings. The source code is public available at https://github.com/zbzhzhy/Unsupervised-Spectral-Reconstruction.
****
Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization
Linjiang Huang, Liang Wang, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8002-8011
As a challenging task of high-level video understanding, weakly supervised temporal action localization has been attracting increasing attention. With only video annotations, most existing methods seek to handle this task with a localization-by-classification framework, which generally adopts a selector to select snippets of high probabilities of actions or namely the foreground. Nevertheless, the existing foreground selection strategies have a major limitation of only considering the unilateral relation from foreground to actions, which cannot guarantee the foreground-action consistency. In this paper, we present a framework named FAC-Net based on the I3D backbone, on which three branches are appended, named class-wise foreground classification branch, class-agnostic attention branch and multiple instance learning branch. First, our class-wise foreground classification branch regularizes the relation between actions and foreground to maximize th

e foreground-background separation. Besides, the class-agnostic attention branch and multiple instance learning branch are adopted to regularize the foreground-action consistency and help to learn a meaningful foreground classifier. Within each branch, we introduce a hybrid attention mechanism, which calculates multiple attention scores for each snippet, to focus on both discriminative and less-discriminative snippets to capture the full action boundaries. Experimental results on THUMOS14 and ActivityNet1.3 demonstrate the superior performance over state-of-the-art approaches.
****

MixMo: Mixing Multiple Inputs for Multiple Outputs via Deep Subnetworks
Alexandre Ramé, Rémy Sun, Matthieu Cord; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 823-833
Recent strategies achieved ensembling ""for free"" by fitting concurrently diverse subnetworks inside a single base network. The main idea during training is that each subnetwork learns to classify only one of the multiple inputs simultaneously provided. However, the question of how to best mix these multiple inputs has not been studied so far. In this paper, we introduce MixMo, a new generalized framework for learning multi-input multi-output deep subnetworks. Our key motivation is to replace the suboptimal summing operation hidden in previous approaches by a more appropriate mixing mechanism. For that purpose, we draw inspiration from successful mixed sample data augmentations. We show that binary mixing in features - particularly with rectangular patches from CutMix - enhances results by making subnetworks stronger and more diverse. We improve state of the art for image classification on CIFAR-100 and Tiny ImageNet datasets. Our easy to implement models notably outperform data augmented deep ensembles, without the inference and memory overheads. As we operate in features and simply better leverage the expressiveness of large networks, we open a new line of research complementary to previous works.
****

CrowdDriven: A New Challenging Dataset for Outdoor Visual Localization
Ara Jafarzadeh, Manuel López Antequera, Pau Gargallo, Yubin Kuang, Carl Toft, Fredrik Kahl, Torsten Sattler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9845-9855
Visual localization is the problem of estimating the position and orientation from which a given image (or a sequence of images) is taken in a known scene. It is an important part of a wide range of computer vision and robotics applications, from self-driving cars to augmented/virtual reality systems. Visual localization techniques should work reliably and robustly under a wide range of conditions, including seasonal, weather, illumination and man-made changes. Recent benchmarking efforts model this by providing images under different conditions, and the community has made rapid progress on these datasets since their inception. However, they are limited to a few geographical regions and often recorded with a single device. We propose a new benchmark for visual localization in outdoor scenes, using crowd-sourced data to cover a wide range of geographical regions and camera devices with a focus on the failure cases of current algorithms. Experiments with state-of-the-art localization approaches show that our dataset is very challenging, with all evaluated methods failing on its hardest parts. As part of the dataset release, we provide the tooling used to generate it, enabling efficient and effective 2D correspondence annotation to obtain reference poses.
****

PnP-DETR: Towards Efficient Visual Analysis With Transformers
Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, Shuicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4661-4670
Recently, DETR pioneered the solution of vision tasks with transformers, it directly translates the image feature map into the object detection result. Though effective, translating the full feature map can be costly due to redundant computation on some area like the background. In this work, we encapsulate the idea of reducing spatial redundancy into a novel poll and pool (PnP) sampling module, with which we build an end-to-end PnP-DETR architecture that adaptively allocates its computation spatially to be more efficient. Concretely, the PnP module abst

racts the image feature map into fine foreground object feature vectors and a sm
all number of coarse background contextual feature vectors. The transformer mode
ls information interaction within the fine-coarse feature space and translates t
he features into the detection result. Moreover, the PnP-augmented model can ins
tantly achieve various desired trade-offs between performance and computation wi
th a single model by varying the sampled feature length, without requiring to tr
ain multiple models as existing methods. Thus it offers greater flexibility for
deployment in diverse scenarios with varying computation constraint. We further
validate the generalizability of the PnP module on panoptic segmentation and the
 recent transformer-based image recognition model ViT and show consistent effici
ency gain. We believe our method makes a step for efficient visual analysis with
 transformers, wherein spatial redundancy is commonly observed. Code and models
will be available.
****

PlaneTR: Structure-Guided Transformers for 3D Plane Recovery
Bin Tan, Nan Xue, Song Bai, Tianfu Wu, Gui-Song Xia; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 4186-4195
This paper presents a neural network built upon Transformers, namely PlaneTR, to
 simultaneously detect and reconstruct planes from a single image. Different fro
m previous methods, PlaneTR jointly leverages the context information and the ge
ometric structures in a sequence-to-sequence way to holistically detect plane in
stances in one forward pass. Specifically, we represent the geometric structures
 as line segments and conduct the network with three main components: (i) contex
t and line segments encoders, (ii) a structure-guided plane decoder, (iii) a pix
el-wise plane embedding decoder. Given an image and its detected line segments,
PlaneTR generates the context and line segment sequences via two specially desig
ned encoders and then feeds them into a Transformers-based decoder to directly p
redict a sequence of plane instances by simultaneously considering the context a
nd global structure cues. Finally, the pixel-wise embeddings are computed to ass
ign each pixel to one predicted plane instance which is nearest to it in embeddi
ng space. Comprehensive experiments demonstrate that PlaneTR achieves state-of-t
he-art performance on the ScanNet and NYUv2 datasets.
****

Concept Generalization in Visual Representation Learning
Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, Karteek Alahari; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 9629-9639
Measuring concept generalization, i.e., the extent to which models trained on a
set of (seen) visual concepts can be leveraged to recognize a new set of (unseen
) concepts, is a popular way of evaluating visual representations, especially in
 a self-supervised learning framework. Nonetheless, the choice of unseen concept
s for such an evaluation is usually made arbitrarily, and independently from the
 seen concepts used to train representations, thus ignoring any semantic relatio
nships between the two. In this paper, we argue that the semantic relationships
between seen and unseen concepts affect generalization performance and propose I
mageNet-CoG, a novel benchmark on the ImageNet-21K (IN-21K) dataset that enables
 measuring concept generalization in a principled way. Our benchmark leverages e
xpert knowledge that comes from WordNet in order to define a sequence of unseen
IN-21K concept sets that are semantically more and more distant from the ImageNe
t-1K (IN-1K) subset, a ubiquitous training set. This allows us to benchmark visu
al representations learned on IN-1K out-of-the box. We conduct a large-scale stu
dy encompassing 31 convolution and transformer-based models and show how differe
nt architectures, levels of supervision, regularization techniques and use of we
b data impact the concept generalization performance.
****

Unsupervised Segmentation Incorporating Shape Prior via Generative Adversarial N
etworks
Dahye Kim, Byung-Woo Hong; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2021, pp. 7324-7334
We present an image segmentation algorithm that is developed in an unsupervised

deep learning framework. The delineation of object boundaries often fails due to the nuisance factors such as illumination changes and occlusions. Thus, we initially propose an unsupervised image decomposition algorithm to obtain an intrinsic representation that is robust with respect to undesirable bias fields based on a multiplicative image model. The obtained intrinsic image is subsequently provided to an unsupervised segmentation procedure that is developed based on a piecewise smooth model. The segmentation model is further designed to incorporate a geometric constraint imposed in the generative adversarial network framework where the discrepancy between the distribution of partitioning functions and the distribution of prior shapes is minimized. We demonstrate the effectiveness and robustness of the proposed algorithm in particular with bias fields and occlusions using simple yet illustrative synthetic examples and a benchmark dataset for image segmentation.

****

DRB-GAN: A Dynamic ResBlock Generative Adversarial Network for Artistic Style Transfer

Wenju Xu, Chengjiang Long, Ruisheng Wang, Guanghui Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6383-6392

In this work, we propose a Dynamic ResBlock Generative Adversarial Network (DRB-GAN) for artistic style transfer. The style code is modeled as the shared parameters for Dynamic ResBlocks connecting both the style encoding network and the style transfer network. In the style encoding network, a style class-aware attention mechanism is used to attend the style feature represent for generating the style codes. In the style transfer network, multiple Dynamic ResBlocks are designed to integrate the style code and the extracted CNN semantic feature and and then feed into the spatial window Layer-Instance Normalization (SW-LIN) decoder, which enables high-quality synthetic images with artistic style transfer. Moreover, the style collection conditional discriminator is designed to ensure our DRB-GAN model to equip with abilities for both arbitrary style transfer and collection style transfer during the training stage. No matter for arbitrary style transfer or collection style transfer, extensive experimental results strongly demonstrate that our proposed DRB-GAN beats state-of-the-art methods and exhibits its superior performance in terms of visual quality and efficiency.

****

Act the Part: Learning Interaction Strategies for Articulated Object Part Discovery

Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15752-15761

People often use physical intuition when manipulating articulated objects, irrespective of object semantics. Motivated by this observation, we identify an important embodied task where an agent must play with objects to recover their parts. To this end, we introduce Act the Part (AtP) to learn how to interact with articulated objects to discover and segment their pieces. By coupling action selection and motion segmentation, AtP is able to isolate structures to make perceptual part recovery possible without semantic labels. Our experiments show AtP learns efficient strategies for part discovery, can generalize to unseen categories, and is capable of conditional reasoning for the task. Although trained in simulation, we show convincing transfer to real world data with no fine-tuning. A summery video, interactive demo, and code will be available at atp.cs.columbia.edu.

****

DCT-SNN: Using DCT To Distribute Spatial Information Over Time for Low-Latency Spiking Neural Networks

Isha Garg, Sayeed Shafayet Chowdhury, Kaushik Roy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4671-4680

Spiking Neural Networks (SNNs) offer a promising alternative to traditional deep learning frameworks, since they provide higher computational efficiency due to event-driven information processing. SNNs distribute the analog values of pixel intensities into binary spikes over time. However, the most widely used input coding schemes, such as Poisson based rate-coding, do not leverage the additional temporal learning capability of SNNs effectively. Moreover, these SNNs suffer fr

om high inference latency which is a major bottleneck to their deployment. To ov ercome this, we propose a time-based encoding scheme that utilizes the Discrete Cosine Transform (DCT) to reduce the number of timesteps required for inference. DCT decomposes an image into a weighted sum of sinusoidal basis images. At each time step, a single frequency base, taken in order and modulated by its corresp onding DCT coefficient, is input to an accumulator that generates spikes upon cr ossing a threshold. We use the proposed scheme to learn DCT-SNN, a low-latency d eep SNN with leaky-integrate-and-fire neurons, trained using surrogate gradient descent based backpropagation. We achieve top-1 accuracy of 89.94%, 68.3% and 52 .43% on CIFAR-10, CIFAR-100 and TinyImageNet, respectively using VGG architectur es. Notably, DCT-SNN performs inference with 2-14X reduced latency compared to o ther state-of-the-art SNNs, while achieving comparable accuracy to their standar d deep learning counterparts. The dimension of the transform allows us to contro l the number of timesteps required for inference. Additionally, we can trade-off accuracy with latency in a principled manner by dropping the highest frequency components during inference. The code is publicly available*.

****

Learning To Resize Images for Computer Vision Tasks

Hossein Talebi, Peyman Milanfar; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 497-506

For all the ways convolutional neural nets have revolutionized computer vision i n recent years, one important aspect has received surprisingly little attention: the effect of image size on the accuracy of tasks being trained for. Typically, to be efficient, the input images are resized to a relatively small spatial res olution (e.g. 224x224), and both training and inference are carried out at this resolution. The actual mechanism for this re-scaling has been an afterthought: N amely, off-the-shelf image resizers such as bilinear and bicubic are commonly us ed in most machine learning software frameworks. But do these resizers limit the on task performance of the trained networks? The answer is yes. Indeed, we show that the typical linear resizer can be replaced with learned resizers that can substantially improve performance. Importantly, while the classical resizers typ ically result in better perceptual quality of the downscaled images, our propose d learned resizers do not necessarily give better visual quality, but instead im prove task performance. Our learned image resizer is jointly trained with a base line vision model. This learned CNN-based resizer creates machine friendly visua l manipulations that lead to a consistent improvement of the end task metric ove r the baseline model. Specifically, here we focus on the classification task wit h the ImageNet dataset, and experiment with four different models to learn resiz ers adapted to each model. Moreover, we show that the proposed resizer can also be useful for fine-tuning the classification baselines for other vision tasks. T o this end, we experiment with three different baselines to develop image qualit y assessment (IQA) models on the AVA dataset.

****

Self-Supervised Cryo-Electron Tomography Volumetric Image Restoration From Singl e Noisy Volume With Sparsity Constraint

Zhidong Yang, Fa Zhang, Renmin Han; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2021, pp. 4056-4065

Cryo-Electron Tomography (cryo-ET) is a powerful tool for 3D cellular visualizat ion. Due to instrumental limitations, cryo-ET images and their volumetric recons truction suffer from extremely low signal-to-noise ratio. In this paper, we prop ose a novel end-to-end self-supervised learning model, the Sparsity Constrained Network (SC-Net), to restore volumetric image from single noisy data in cryo-ET. The proposed method only requires a single noisy data as training input and no ground-truth is needed in the whole training procedure. A new target function is proposed to preserve both local smoothness and detailed structure. Additionally , a novel procedure for the simulation of electron tomographic photographing is designed to help the evaluation of methods. Experiments are done on three simula ted data and four real-world data. The results show that our method could produc e a strong enhancement for a single very noisy cryo-ET volumetric data, which is much better than the state-of-the-art Noise2Void, and with a competitive perfor

mance comparing with Noise2Noise. Code is available at https://github.com/icthrm/SC-Net.
****

The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization
Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, Justin Gilmer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8340-8349
We introduce four new real-world distribution shift datasets consisting of changes in image style, image blurriness, geographic location, camera operation, and more. With our new datasets, we take stock of previously proposed methods for improving out-of-distribution robustness and put them to the test. We find that using larger models and artificial data augmentations can improve robustness on real-world distribution shifts, contrary to claims in prior work. We find improvements in artificial robustness benchmarks can transfer to real-world distribution shifts, contrary to claims in prior work. Motivated by our observation that data augmentations can help with real-world distribution shifts, we also introduce a new data augmentation method which advances the state-of-the-art and outperforms models pretrained with 1000x more labeled data. Overall we find that some methods consistently help with distribution shifts in texture and local image statistics, but these methods do not help with some other distribution shifts like geographic changes. Our results show that future research must study multiple distribution shifts simultaneously, as we demonstrate that no evaluated method consistently improves robustness.
****

Field of Junctions: Extracting Boundary Structure at Low SNR
Dor Verbin, Todd Zickler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6869-6878
We introduce a bottom-up model for simultaneously finding many boundary elements in an image, including contours, corners and junctions. The model explains boundary shape in each small patch using a 'generalized M-junction' comprising M angles and a freely-moving vertex. Images are analyzed using non-convex optimization to cooperatively find M+2 junction values at every location, with spatial consistency being enforced by a novel regularizer that reduces curvature while preserving corners and junctions. The resulting 'field of junctions' is simultaneously a contour detector, corner/junction detector, and boundary-aware smoothing of regional appearance. Notably, its unified analysis of contours, corners, junctions and uniform regions allows it to succeed at high noise levels, where other methods for segmentation and boundary detection fail.
****

Crowd Counting With Partial Annotations in an Image
Yanyu Xu, Ziming Zhong, Dongze Lian, Jing Li, Zhengxin Li, Xinxing Xu, Shenghua Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15570-15579
To fully leverage the data captured from different scenes with different view angles while reducing the annotation cost, this paper studies a novel crowd counting setting, i.e. only using partial annotations in each image as training data. Inspired by the repetitive patterns in the annotated and unannotated regions as well as the ones between them, we design a network with three components to tackle those unannotated regions: i) in an Unannotated Regions Characterization (URC) module, we employ a memory bank to only store the annotated features, which could help the visual features extracted from these annotated regions flow to these unannotated regions; ii) For each image, Feature Distribution Consistency (FDC) regularizes the feature distributions of annotated head and unannotated head regions to be consistent; iii) a Cross-regressor Consistency Regularization (CCR) module is designed to learn the visual features of unannotated regions in a self-supervised style. The experimental results validate the effectiveness of our proposed model under the partial annotation setting for several datasets, such as ShanghaiTech, UCF-CC-50, UCF-QNRF, NWPU-Crowd, and JHU-CROWD++. With only 10% a

nnotated regions in each image, our proposed model achieves better performance t
han the recent methods and baselines under semi-supervised or active learning se
ttings on all datasets. The code is https://github.com/svip-lab/CrwodCountingPAL
.
****
Continual Neural Mapping: Learning an Implicit Scene Representation From Sequent
ial Observations
Zike Yan, Yuxin Tian, Xuesong Shi, Ping Guo, Peng Wang, Hongbin Zha; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1
5782-15792
Recent advances have enabled a single neural network to serve as an implicit sce
ne representation, establishing the mapping function between spatial coordinates
 and scene properties. In this paper, we make a further step towards continual l
earning of the implicit scene representation directly from sequential observatio
ns, namely Continual Neural Mapping. The proposed problem setting bridges the ga
p between batch-trained implicit neural representations and commonly used stream
ing data in robotics and vision communities. We introduce an experience replay a
pproach to tackle an exemplary task of continual neural mapping: approximating a
 continuous signed distance function (SDF) from sequential depth images as a sce
ne geometry representation. We show for the first time that a single network can
 represent scene geometry over time continually without catastrophic forgetting,
 while achieving promising trade-offs between accuracy and efficiency.
****
LSD-StructureNet: Modeling Levels of Structural Detail in 3D Part Hierarchies
Dominic Roberts, Ara Danielyan, Hang Chu, Mani Golparvar-Fard, David Forsyth; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
21, pp. 5836-5845
Generative models for 3D shapes represented by hierarchies of parts can generate
 realistic and diverse sets of outputs. However, existing models suffer from the
 key practical limitation of modelling shapes holistically and thus cannot perfo
rm conditional sampling, i.e. they are not able to generate variants on individu
al parts of generated shapes without modifying the rest of the shape. This is li
miting for applications such as 3D CAD design that involve adjusting created sha
pes at multiple levels of detail. To address this, we introduce LSD-StructureNet
, an augmentation to the StructureNet architecture that enables re-generation of
 parts situated at arbitrary positions in the hierarchies of its outputs. We ach
ieve this by learning individual, probabilistic conditional decoders for each hi
erarchy depth. We evaluate LSD-StructureNet on the PartNet dataset, the largest
dataset of 3D shapes represented by hierarchies of parts. Our results show that
contrarily to existing methods, LSD-StructureNet can perform conditional samplin
g without impacting inference speed or the realism and diversity of its outputs.
****
Weakly Supervised Temporal Anomaly Segmentation With Dynamic Time Warping
Dongha Lee, Sehun Yu, Hyunjun Ju, Hwanjo Yu; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2021, pp. 7355-7364
Most recent studies on detecting and localizing temporal anomalies have mainly e
mployed deep neural networks to learn the normal patterns of temporal data in an
 unsupervised manner. Unlike them, the goal of our work is to fully utilize inst
ance-level (or weak) anomaly labels, which only indicate whether any anomalous e
vents occurred or not in each instance of temporal data. In this paper, we prese
nt WETAS, a novel framework that effectively identifies anomalous temporal segme
nts (i.e., consecutive time points) in an input instance. WETAS learns discrimin
ative features from the instance-level labels so that it infers the sequential o
rder of normal and anomalous segments within each instance, which can be used as
 a rough segmentation mask. Based on the dynamic time warping (DTW) alignment be
tween the input instance and its segmentation mask, WETAS obtains the result of
temporal segmentation, and simultaneously, it further enhances itself by using t
he mask as additional supervision. Our experiments show that WETAS considerably
outperforms other baselines in terms of the localization of temporal anomalies,
and also it provides more informative results than point-level detection methods

.
****
Adaptive Label Noise Cleaning With Meta-Supervision for Deep Face Recognition
Yaobin Zhang, Weihong Deng, Yaoyao Zhong, Jiani Hu, Xian Li, Dongyue Zhao, Dongchao Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15065-15075
The training of a deep face recognition system usually faces the interference of label noise in the training data. However, it is difficult to obtain a high-precision cleaning model to remove these noises. In this paper, we propose an adaptive label noise cleaning algorithm based on meta-learning for face recognition datasets, which can learn the distribution of the data to be cleaned and make automatic adjustments based on class differences. It first learns reliable cleaning knowledge from well-labeled noisy data, then gradually transfers it to the target data with meta-supervision to improve performance. A threshold adapter module is also proposed to address the drift problem in transfer learning methods. Extensive experiments clean two noisy in-the-wild face recognition datasets and show the effectiveness of the proposed method to reach state-of-the-art performance on the IJB-C face recognition benchmark.
****
Unsupervised Dense Deformation Embedding Network for Template-Free Shape Correspondence
Ronghan Chen, Yang Cong, Jiahua Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8361-8370
Shape correspondence from 3D deformation learning has attracted appealing academy interests recently. Nevertheless, current deep learning based methods require the supervision of dense annotations to learn per-point translations, which severely over-parameterize the deformation process. Moreover, they fail to capture local geometric details of original shape via global feature embedding. To address these challenges, we develop a new Unsupervised Dense Deformation Embedding Network (i.e., UD2E-Net), which learns to predict deformations between non-rigid shapes from dense local features. Since it is non-trivial to match deformation-variant local features for deformation prediction, we develop an Extrinsic-Intrinsic Autoencoder to frst encode extrinsic geometric features from source into intrinsic coordinates in a shared canonical shape, with which the decoder then synthesizes corresponding target features. Moreover, a bounded maximum mean discrepancy loss is developed to mitigate the distribution divergence between the synthesized and original features. To learn natural deformation without dense supervision, we introduce a coarse parameterized deformation graph, for which a novel trace and propagation algorithm is proposed to improve both the quality and effciency of the deformation. Our UD2E-Net outperforms state-of-the-art unsupervised methods by 24% on Faust Inter challenge and even supervised methods by 13% on Faust Intra challenge.
****
Learning Action Completeness From Points for Weakly-Supervised Temporal Action Localization
Pilhyeon Lee, Hyeran Byun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13648-13657
We tackle the problem of localizing temporal intervals of actions with only a single frame label for each action instance for training. Owing to label sparsity, existing work fails to learn action completeness, resulting in fragmentary action predictions. In this paper, we propose a novel framework, where dense pseudo-labels are generated to provide completeness guidance for the model. Concretely, we first select pseudo background points to supplement point-level action labels. Then, by taking the points as seeds, we search for the optimal sequence that is likely to contain complete action instances while agreeing with the seeds. To learn completeness from the obtained sequence, we introduce two novel losses that contrast action instances with background ones in terms of action score and feature similarity, respectively. Experimental results demonstrate that our completeness guidance indeed helps the model to locate complete action instances, leading to large performance gains especially under high IoU thresholds. Moreover,

we demonstrate the superiority of our method over existing state-of-the-art methods on four benchmarks: THUMOS'14, GTEA, BEOID, and ActivityNet. Notably, our method even performs comparably to recent fully-supervised methods, at the 6 times cheaper annotation cost. Our code is available at https://github.com/Pilhyeon.
****

Re-Distributing Biased Pseudo Labels for Semi-Supervised Semantic Segmentation: A Baseline Investigation
Ruifei He, Jihan Yang, Xiaojuan Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6930-6940
While self-training has advanced semi-supervised semantic segmentation, it severely suffers from the long-tailed class distribution on real-world semantic segmentation datasets that make the pseudo-labeled data bias toward majority classes. In this paper, we present a simple and yet effective Distribution Alignment and Random Sampling (DARS) method to produce unbiased pseudo labels that match the true class distribution estimated from the labeled data. Besides, we also contribute a progressive data augmentation and labeling strategy to facilitate model training with pseudo-labeled data. Experiments on both Cityscapes and PASCAL VOC 2012 datasets demonstrate the effectiveness of our approach. Albeit simple, our method performs favorably in comparison with state-of-the-art approaches. Code will be available at https://github.com/CVMI-Lab/DARS.
****

Visformer: The Vision-Friendly Transformer
Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 589-598
The past year has witnessed the rapid development of applying the Transformer module to vision problems. While some researchers have demonstrated that Transformer-based models enjoy a favorable ability of fitting data, there are still growing number of evidences showing that these models suffer over-fitting especially when the training data is limited. This paper offers an empirical study by performing step-by-step operations to gradually transit a Transformer-based model to a convolution-based model. The results we obtain during the transition process deliver useful messages for improving visual recognition. Based on these observations, we propose a new architecture named Visformer, which is abbreviated from the 'Vision-friendly Transformer'. With the same computational complexity, Visformer outperforms both the Transformer-based and convolution-based models in terms of ImageNet classification accuracy, and the advantage becomes more significant when the model complexity is lower or the training set is smaller. The code is available at https://github.com/danczs/Visformer.
****

Learning Indoor Inverse Rendering With 3D Spatially-Varying Lighting
Zian Wang, Jonah Philion, Sanja Fidler, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12538-12547
In this work, we address the problem of jointly estimating albedo, normals, depth and 3D spatially-varying lighting from a single image. Most existing methods formulate the task as image-to-image translation, ignoring the 3D properties of the scene. However, indoor scenes contain complex 3D light transport where a 2D representation is insufficient. In this paper, we propose a unified, learning-based inverse rendering framework that formulates 3D spatially-varying lighting. Inspired by classic volume rendering techniques, we propose a novel Volumetric Spherical Gaussian representation for lighting, which parameterizes the exitant radiance of the 3D scene surfaces on a voxel grid. We design a physicsbased differentiable renderer that utilizes our 3D lighting representation, and formulates the energy-conserving image formation process that enables joint training of all intrinsic properties with the re-rendering constraint. Our model ensures physically correct predictions and avoids the need for ground-truth HDR lighting which is not easily accessible. Experiments show that our method outperforms prior works both quantitatively and qualitatively, and is capable of producing photorealistic results for AR applications such as virtual object insertion even for highly specular objects.

****
DeepGaze IIE: Calibrated Prediction in and Out-of-Domain for State-of-the-Art Saliency Modeling
Akis Linardos, Matthias Kümmerer, Ori Press, Matthias Bethge; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12919-12928
Since 2014 transfer learning has become the key driver for the improvement of spatial saliency prediction - however, with stagnant progress in the last 3-5 years. We conduct a large-scale transfer learning study which tests different ImageNet backbones, always using the same read out architecture and learning protocol adopted from DeepGaze II. By replacing the VGG19 backbone of DeepGaze II with ResNet50 features we improve the performance on saliency prediction from 78% to 85%. However, as we continue to test better ImageNet models as backbones - such as EfficientNetB5 - we observe no additional improvement on saliency prediction. By analyzing the backbones further, we find that generalization to other datasets differs substantially, with models being consistently overconfident in their fixation predictions. We show that by combining multiple backbones in a principled manner a good confidence calibration on unseen datasets can be achieved. This new model "DeepGaze IIE" yields a significant leap in benchmark performance in and out-of-domain with a 15 percent point improvement over DeepGaze II to 93% on MIT1003, marking a new state of the art on the MIT/Tuebingen Saliency Benchmark in all available metrics (AUC: 88.3%, sAUC: 79.4%, CC: 82.4%).
****
Learning To Drive From a World on Rails
Dian Chen, Vladlen Koltun, Philipp Krähenbühl; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15590-15599
We learn an interactive vision-based driving policy from pre-recorded driving logs via a model-based approach. A forward model of the world supervises a driving policy that predicts the outcome of any potential driving trajectory. To support learning from pre-recorded logs, we assume that the world is on rails, meaning neither the agent nor its actions influence the environment. This assumption greatly simplifies the learning problem, factorizing the dynamics into a nonreactive world model and a low-dimensional and compact forward model of the ego-vehicle. Our approach computes action-values for each training trajectory using a tabular dynamic-programming evaluation of the Bellman equations; these action-values in turn supervise the final vision-based driving policy. Despite the world-on-rails assumption, the final driving policy acts well in a dynamic and reactive world. It outperforms imitation learning as well as model-based and model-free reinforcement learning on the challenging CARLA NoCrash benchmark. It is also an order of magnitude more sample-efficient than state-of-the-art model-free reinforcement learning techniques on navigational tasks in the ProcGen benchmark.
****
Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies
Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, Hujun Bao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14314-14323
This paper addresses the challenge of reconstructing an animatable human model from a multi-view video. Some recent works have proposed to decompose a non-rigidly deforming scene into a canonical neural radiance field and a set of deformation fields that map observation-space points to the canonical space, thereby enabling them to learn the dynamic scene from images. However, they represent the deformation field as translational vector field or SE(3) field, which makes the optimization highly under-constrained. Moreover, these representations cannot be explicitly controlled by input motions. Instead, we introduce neural blend weight fields to produce the deformation fields. Based on the skeleton-driven deformation, blend weight fields are used with 3D human skeletons to generate observation-to-canonical and canonical-to-observation correspondences. Since 3D human skeletons are more observable, they can regularize the learning of deformation fields. Moreover, the learned blend weight fields can be combined with input skeletal motions to generate new deformation fields to animate the human model. Experime

nts show that our approach significantly outperforms recent human synthesis meth
ods. The code and supplementary materials are available at \href https://zju3dv.
github.io/animatable_nerf/ https://zju3dv.github.io/animatable_nerf/ .
****

OpenGAN: Open-Set Recognition via Open Data Generation
Shu Kong, Deva Ramanan; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2021, pp. 813-822
Real-world machine learning systems need to analyze novel testing data that diff
ers from the training data. In K-way classification, this is crisply formulated
as open-set recognition, core to which is the ability to discriminate open-set d
ata outside the K closed-set classes. Two conceptually elegant ideas for open-se
t discrimination are: 1) discriminatively learning an open-vs-closed binary disc
riminator by exploiting some outlier data as the open-set, and 2) unsupervised l
earning the closed-set data distribution with a GAN and using its discriminator
as the open-set likelihood function. However, the former generalizes poorly to d
iverse open test data due to overfitting to the training outliers, which unlikel
y exhaustively span the open-world. The latter does not work well, presumably du
e to the instable training of GANs. Motivated by the above, we propose OpenGAN,
which addresses the limitation of each approach by combining them with several t
echnical insights. First, we show that a carefully selected GAN-discriminator on
 some real outlier data already achieves the state-of-the-art. Second, we augmen
t the available set of real open training examples with adversarially synthesize
d ""fake"" data. Third and most importantly, we build the discriminator over the
 features computed by the closed-world K-way networks. Extensive experiments sho
w that OpenGAN significantly outperforms prior open-set methods.
****

Learning To Reduce Defocus Blur by Realistically Modeling Dual-Pixel Data
Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S. Brown, Peyman Mi
lanfar; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 2289-2298
Recent work has shown impressive results on data-driven defocus deblurring using
 the two-image views available on modern dual-pixel (DP) sensors. One significan
t challenge in this line of research is access to DP data. Despite many cameras
having DP sensors, only a limited number provide access to the low-level DP sens
or images. In addition, capturing training data for defocus deblurring involves
a time-consuming and tedious setup requiring the camera's aperture to be adjuste
d. Some cameras with DP sensors (e.g., smartphones) do not have adjustable apert
ures, further limiting the ability to produce the necessary training data. We ad
dress the data capture bottleneck by proposing a procedure to generate realistic
 DP data synthetically. Our synthesis approach mimics the optical image formatio
n found on DP sensors and can be applied to virtual scenes rendered with standar
d computer software. Leveraging these realistic synthetic DP images, we introduc
e a recurrent convolutional network (RCN) architecture that improves deblurring
results and is suitable for use with single-frame and multi-frame data (e.g., vi
deo) captured by DP sensors. Finally, we show that our synthetic DP data is usef
ul for training DNN models targeting video deblurring applications where access
to DP data remains challenging.
****

Uncertainty-Aware Pseudo Label Refinery for Domain Adaptive Semantic Segmentatio
n
Yuxi Wang, Junran Peng, ZhaoXiang Zhang; Proceedings of the IEEE/CVF Internation
al Conference on Computer Vision (ICCV), 2021, pp. 9092-9101
Unsupervised domain adaptation for semantic segmentation aims to assign the pixe
l-level labels for unlabeled target domain by transferring knowledge from the la
beled source domain. A typical self-supervised learning approach generates pseud
o labels from the source model and then re-trains the model to fit the target di
stribution. However, it suffers from noisy pseudo labels due to the existence of
 domain shift. Related works alleviate this problem by selecting high-confidence
 predictions, but uncertain classes with low confidence scores have rarely been
considered. This informative uncertainty is essential to enhance feature represe

ntation and align source and target domains. In this paper, we propose a novel u ncertainty-aware pseudo label refinery framework considering two crucial factors simultaneously. First, we progressively enhance the feature alignment model via the target-guided uncertainty rectifying framework. Second, we provide an uncer tainty-aware pseudo label assignment strategy without any manually designed thre shold to reduce the noisy labels. Extensive experiments demonstrate the effectiv eness of our proposed approach and achieve state-of-the-art performance on two s tandard synthetic-2-real tasks.
****

Mining Contextual Information Beyond Image for Semantic Segmentation
Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, Jie Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7231-7241
This paper studies the context aggregation problem in semantic image segmentatio n. The existing researches focus on improving the pixel representations by aggre gating the contextual information within individual images. Though impressive, t hese methods neglect the significance of the representations of the pixels of th e corresponding class beyond the input image. To address this, this paper propos es to mine the contextual information beyond individual images to further augmen t the pixel representations. We first set up a feature memory module, which is u pdated dynamically during training, to store the dataset-level representations o f various categories. Then, we learn class probability distribution of each pixe l representation under the supervision of the ground-truth segmentation. At last , the representation of each pixel is augmented by aggregating the dataset-level representations based on the corresponding class probability distribution. Furt hermore, by utilizing the stored dataset-level representations, we also propose a representation consistent learning strategy to make the classification head be tter address intra-class compactness and inter-class dispersion. The proposed me thod could be effortlessly incorporated into existing segmentation frameworks (e .g., FCN, PSPNet, OCRNet and DeepLabV3) and brings consistent performance improv ements. Mining contextual information beyond image allows us to report state-of- the-art performance on various benchmarks: ADE20K, LIP, Cityscapes and COCO-Stuf f.
****

A General Recurrent Tracking Framework Without Real Data
Shuai Wang, Hao Sheng, Yang Zhang, Yubin Wu, Zhang Xiong; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13219-13228
Recent progress in multi-object tracking (MOT) has shown great significance of a robust scoring mechanism for potential tracks. However, the lack of available d ata in MOT makes it difficult to learn a general scoring mechanism. Multiple cue s including appearance, motion and etc., are limitedly utilized in current manua l scoring functions. In this paper, we propose a Multiple Nodes Tracking (MNT) f ramework that adapts to most trackers. Based on this framework, a Recurrent Trac king Unit (RTU) is designed to score potential tracks through long-term informat ion. In addition, we present a method of generating simulated tracking data with out real data to overcome the defect of limited available data in MOT. The exper iments demonstrate that our simulated tracking data is effective for training RT U and achieves state-of-the-art performance on both MOT17 and MOT16 benchmarks. Meanwhile, RTU can be flexibly plugged into classic trackers such as DeepSORT an d MHT, and makes remarkable improvements as well.
****

Knowledge Mining and Transferring for Domain Adaptive Object Detection
Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9133 -9142
With the thriving of deep learning, CNN-based object detectors have made great p rogress in the past decade. However, the domain gap between training and testing data leads to a prominent performance degradation and thus hinders their applic ation in the real world. To alleviate this problem, Knowledge Transfer Network (KTNet) is proposed as a new paradigm for domain adaption. Specifically, KTNet is

constructed on a base detector with intrinsic knowledge mining and relational knowledge constraints. First, we design a foreground/background classifier shared by source domain and target domain to extract the common attribute knowledge of objects in different scenarios. Second, we model the relational knowledge graph and explicitly constrain the consistency of category correlation under source domain, target domain, as well as cross-domain conditions. As a result, the detector is guided to learn object-related and domain-independent representation. Extensive experiments and visualizations confirm that transferring object-specific knowledge can yield notable performance gains. The proposed KTNet achieves state-of-the-art results on three cross-domain detection benchmarks.

****

Cloud Transformers: A Universal Approach to Point Cloud Processing Tasks
Kirill Mazur, Victor Lempitsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10715-10724

We present a new versatile building block for deep point cloud processing architectures that is equally suited for diverse tasks. This building block combines the ideas of spatial transformers and multi-view convolutional networks with the efficiency of standard convolutional layers in two and three-dimensional dense grids. The new block operates via multiple parallel heads, whereas each head differentiably rasterizes feature representations of individual points into a low-dimensional space, and then uses dense convolution to propagate information across points. The results of the processing of individual heads are then combined together resulting in the update of point features. Using the new block, we build architectures for both discriminative (point cloud segmentation, point cloud classification) and generative (point cloud inpainting and image-based point cloud reconstruction) tasks. The resulting architectures achieve state-of-the-art performance for these tasks, demonstrating the versatility of the new block for point cloud processing.

****

TravelNet: Self-Supervised Physically Plausible Hand Motion Learning From Monocular Color Images
Zimeng Zhao, Xi Zhao, Yangang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11666-11676

This paper aims to reconstruct physically plausible hand motion from monocular color images. Existing frame-by-frame estimating approaches can not guarantee the physical plausibility (e.g. penetration, jittering) directly. In this paper, we embed physical constraints on the per-frame estimated motions in both spatial and temporal space. Our key idea is to adopt a self-supervised learning strategy to train a novel encoder-decoder, named TravelNet, whose training motion data is prepared by the physics engine using discrete pose states. TravelNet captures key pose states from hand motion sequences as compact motion descriptors, inspired by the concept of keyframes in animation. Finally, it manages to extract those key states out of perturbations without manual annotations, and reconstruct the motions preserving details and physical plausibility. In the experiments, we show that the outputs of the TravelNet contain both finger synergism and time consistency. Through the proposed framework, hand motions can be accurately reconstructed and flexibly re-edited, which is superior to the state-of-the-art methods.

****

Joint Visual Semantic Reasoning: Multi-Stage Decoder for Text Recognition
Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, Yi-Zhe Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14940-14949

Although text recognition has significantly evolved over the years, state-of the-art (SOTA) models still struggle in the wild scenarios due to complex backgrounds, varying fonts, uncontrolled illuminations, distortions and other artifacts. This is because such models solely depend on visual information for text recognition, thus lacking semantic reasoning capabilities. In this paper, we argue that semantic information offers a complementary role in addition to visual only. More specifically, we additionally utilize semantic information by proposing a multi-stage multi-scale attentional decoder that performs joint visual-semantic rea

soning. Our novelty lies in the intuition that for text recognition, prediction should be refined in a stage-wise manner. Therefore our key contribution is in designing a stage-wise unrolling attentional decoder where non-differentiability, invoked by discretely predicted character labels, needs to be bypassed for end-to-end training. While the first stage predicts using visual features, subsequent stages refine on-top of it using joint visual-semantic information. Additionally, we introduce multi-scale 2D attention along with dense and residual connections between different stages to deal with varying scales of character sizes, for better performance and faster convergence during training. Experimental results show our approach to outperform existing SOTA methods by a considerable margin.
****

Video Self-Stitching Graph Network for Temporal Action Localization
Chen Zhao, Ali K. Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13658-13667
Temporal action localization (TAL) in videos is a challenging task, especially due to the large variation in action temporal scales. Short actions usually occupy a major proportion in the datasets, but tend to have the lowest performance. In this paper, we confront the challenge of short actions and propose a multi-level cross-scale solution dubbed as video self-stitching graph network (VSGN). We have two key components in VSGN: video self-stitching (VSS) and cross-scale graph pyramid network (xGPN). In VSS, we focus on a short period of a video and magnify it along the temporal dimension to obtain a larger scale. We stitch the original clip and its magnified counterpart in one input sequence to take advantage of the complementary properties of both scales. The xGPN component further exploits the cross-scale correlations by a pyramid of cross-scale graph networks, each containing a hybrid module to aggregate features from across scales as well as within the same scale. Our VSGN not only enhances the feature representations, but also generates more positive anchors for short actions and more short training samples. Experiments demonstrate that VSGN obviously improves the localization performance of short actions as well as achieving the state-of-the-art overall performance on THUMOS-14 and ActivityNet-v1.3.
****

Dynamic Dual Gating Neural Networks
Fanrong Li, Gang Li, Xiangyu He, Jian Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5330-5339
In dynamic neural networks that adapt computations to different inputs, gating-based methods have demonstrated notable generality and applicability in trading-off the model complexity and accuracy. However, existing works only explore the redundancy from a single point of the network, limiting the performance. In this paper, we propose dual gating, a new dynamic computing method, to reduce the model complexity at run-time. For each convolutional block, dual gating identifies the informative features along two separate dimensions, spatial and channel. Specifically, the spatial gating module estimates which areas are essential, and the channel gating module predicts the salient channels that contribute more to the results. Then the computation of both unimportant regions and irrelevant channels can be skipped dynamically during inference. Extensive experiments on a variety of datasets demonstrate that our method can achieve higher accuracy under similar computing budgets compared with other dynamic execution methods. In particular, dynamic dual gating can provide 59.7% saving in computing of ResNet50 with 76.41% top-1 accuracy on ImageNet, which has advanced the state-of-the-art.
****

Gravity-Aware Monocular 3D Human-Object Reconstruction
Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, Vladislav Golyanik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12365-12374
This paper proposes GraviCap, i.e., a new approach for joint markerless 3D human motion capture and object trajectory estimation from monocular RGB videos. We focus on scenes with objects partially observed during a free flight. In contrast to existing monocular methods, we can recover scale, object trajectories as well as human bone lengths in meters and the ground plane's orientation, thanks to

the awareness of the gravity constraining object motions. Our objective function is parametrised by the object's initial velocity and position, gravity direction and focal length, and jointly optimised for one or several free flight episodes. The proposed human-object interaction constraints ensure geometric consistency of the 3D reconstructions and improved physical plausibility of human poses compared to the unconstrained case. We evaluate GraviCap on a new dataset with ground-truth annotations for persons and different objects undergoing free flights. In the experiments, our approach achieves state-of-the-art accuracy in 3D human motion capture on various metrics. We urge the reader to watch our supplementary video. Both the source code and the dataset are released; see http://4dqv.mpi-inf.mpg.de/GraviCap/.

****

Exploring Relational Context for Multi-Task Dense Prediction

David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15869-15878

The timeline of computer vision research is marked with advances in learning and utilizing efficient contextual representations. Most of them, however, are targeted at improving model performance on a single downstream task. We consider a multi-task environment for dense prediction tasks, represented by a common backbone and independent task-specific heads. Our goal is to find the most efficient way to refine each task prediction by capturing cross-task contexts dependent on tasks' relations. We explore various attention-based contexts, such as global and local, in the multi-task setting and analyze their behavior when applied to refine each task independently. Empirical findings confirm that different source-target task pairs benefit from different context types. To automate the selection process, we propose an Adaptive Task-Relational Context (ATRC) module, which samples the pool of all available contexts for each task pair using neural architecture search and outputs the optimal configuration for deployment. Our method achieves state-of-the-art performance on two important multi-task benchmarks, namely NYUD-v2 and PASCAL-Context. The proposed ATRC has a low computational toll and can be used as a drop-in refinement module for any supervised multi-task architecture.

****

Going Deeper With Image Transformers

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, Hervé Jégou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 32-42

Transformers have been recently adapted for large scale image classification, achieving high scores shaking up the long supremacy of convolutional neural networks. However the optimization of vision transformers has been little studied so far. In this work, we build and optimize deeper transformer networks for image classification. In particular, we investigate the interplay of architecture and optimization of such dedicated transformers. We make two architecture changes that significantly improve the accuracy of deep transformers. This leads us to produce models whose performance does not saturate early with more depth, for instance we obtain 86.5% top-1 accuracy on Imagenet when training with no external data, we thus attain the current sate of the art with less floating-point operations and parameters. Our best model establishes the new state of the art on Imagenet with Reassessed labels and Imagenet-V2 / match frequency, in the setting with no additional training data. We share our code and models

****

UltraPose: Synthesizing Dense Pose With 1 Billion Points by Human-Body Decoupling 3D Model

Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, Tianxiang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10891-10900

Recovering dense human poses from images plays a critical role in establishing an image-to-surface correspondence between RGB images and the 3D surface of the human body, serving the foundation of rich real-world applications, such as virtu

al humans, monocular-to-3d reconstruction. However, the popular DensePose-COCO d
ataset relies on a sophisticated manual annotation system, leading to severe lim
itations in acquiring the denser and more accurate annotated pose resources. In
this work, we introduce a new 3D human-body model with a series of decoupled par
ameters that could freely control the generation of the body. Furthermore, we bu
ild a data generation system based on this decoupling 3D model, and construct an
 ultra dense synthetic benchmark UltraPose, containing around 1.3 billion corres
ponding points. Compared to the existing manually annotated DensePose-COCO datas
et, the synthetic UltraPose has ultra dense image-to-surface correspondences wit
hout annotation cost and error. Our proposed UltraPose provides the largest benc
hmark and data resources for lifting the model capability in predicting more acc
urate dense poses. To promote future researches in this field, we also propose a
 transformer-based method to model the dense correspondence between 2D and 3D wo
rlds. The proposed model trained on synthetic UltraPose can be applied to real-w
orld scenarios, indicating the effectiveness of our benchmark and model.
****

Hand Image Understanding via Deep Multi-Task Learning
Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng,
 Lei Wang, Ji Liu; Proceedings of the IEEE/CVF International Conference on Compu
ter Vision (ICCV), 2021, pp. 11281-11292
Analyzing and understanding hand information from multimedia materials like imag
es or videos is important for many real world applications and remains to be ver
y active in research community. There are various works focusing on recovering h
and information from single image, however, they usually solve a single task, fo
r example, hand mask segmentation, 2D/3D hand pose estimation, or hand mesh reco
nstruction and perform not well in challenging scenarios. To further improve the
 performance of these tasks, we propose a novel Hand Image Understanding (HIU) f
ramework (HIU-DMTL) to extract comprehensive information of the hand object from
 a single RGB image, by jointly considering the relationships between these task
s. To achieve this goal, a cascaded multi-task learning (MTL) backbone is design
ed to estimate the 2D heat maps, to learn the segmentation mask, and to generate
 the intermediate 3D information encoding, followed by a coarse-to-fine learning
 paradigm and a self-supervised learning strategy. Qualitative experiments demon
strate that our approach is capable of recovering reasonable mesh representation
s even in challenging situations. Quantitatively, our method significantly outpe
rforms the state-of-the-art approaches on various widely-used datasets, in terms
 of diverse evaluation metrics.
****

MFNet: Multi-Filter Directive Network for Weakly Supervised Salient Object Detec
tion
Yongri Piao, Jian Wang, Miao Zhang, Huchuan Lu; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2021, pp. 4136-4145
Weakly supervised salient object detection (WSOD) targets to train a CNNs-based
saliency network using only low-cost annotations. Existing WSOD methods take var
ious techniques to pursue single "high-quality" pseudo label from low-cost annot
ations and then develop their saliency networks. Though these methods have achie
ved good performance, the generated single label is inevitably affected by adopt
ed refinement algorithms and shows prejudiced characteristics which further infl
uence the saliency networks. In this work, we introduce a new multiple-pseudo la
bel framework to integrate more comprehensive and accurate saliency cues from mu
ltiple labels, avoiding the aforementioned problem. Specifically, we propose a m
ulti-filer directive network (MFNet) including a saliency network as well as mul
tiple directive filters. The directive filter (DF) is designed to extract and fi
lter more accurate saliency cues from the noisy pseudo labels. The multiple accu
rate cues from multiple DFs are then simultaneously propagated to the saliency n
etwork with a multi-guidance loss. Extensive experiments on five datasets over f
our metrics demonstrate that our method outperforms all the existing congeneric
methods. Moreover, it is also worth noting that our framework is flexible enough
 to apply to existing methods and improve their performance.
****

DRIVE: Deep Reinforced Accident Anticipation With Visual Explanation
Wentao Bao, Qi Yu, Yu Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7619-7628
Traffic accident anticipation aims to accurately and promptly predict the occurrence of a future accident from dashcam videos, which is vital for a safety-guaranteed self-driving system. To encourage an early and accurate decision, existing approaches typically focus on capturing the cues of spatial and temporal context before a future accident occurs. However, their decision-making lacks visual explanation and ignores the dynamic interaction with the environment. In this paper, we propose Deep ReInforced accident anticipation with Visual Explanation, named DRIVE. The method simulates both the bottom-up and top-down visual attention mechanism in a dashcam observation environment so that the decision from the proposed stochastic multi-task agent can be visually explained by attentive regions. Moreover, the proposed dense anticipation reward and sparse fixation reward are effective in training the DRIVE model with our improved reinforcement learning algorithm. Experimental results show that the DRIVE model achieves state-of-the-art performance on multiple real-world traffic accident datasets. Code and pre-trained models are available at https://www.rit.edu/actionlab/drive.
****

On the Importance of Distractors for Few-Shot Classification
Rajshekhar Das, Yu-Xiong Wang, José M. F. Moura; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9030-9040
Few-shot classification aims at classifying categories of a novel task by learning from just a few (typically, 1 to 5) labeled examples. An effective approach to few-shot classification involves a prior model trained on a large-sample base domain, which is then finetuned over the novel few-shot task to yield generalizable representations. However, task-specific finetuning is prone to overfitting due to the lack of enough training examples. To alleviate this issue, we propose a new finetuning approach based on contrastive learning that reuses unlabelled examples from the base domain in the form of distractors. Unlike the nature of unlabelled data used in prior works, distractors belong to classes that do not overlap with the novel categories. We demonstrate for the first time that the inclusion of such distractors can significantly boost few-shot generalization. Our technical novelty includes a stochastic pairing of examples sharing the same category in the few-shot task and a weighting term that controls the relative influence of task-specific negatives and distractors. An important aspect of our finetuning objective is that it is agnostic to distractor labels and hence applicable to various base domain settings. More precisely, compared to state-of-the-art approaches, our method shows accuracy gains of up to 12% in cross-domain and up to 5% in unsupervised prior-learning settings. Our code is available at https://github.com/quantacode/Contrastive-Finetuning.git
****

Zero-Shot Natural Language Video Localization
Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1470-1479
Understanding videos to localize moments with natural language often requires large expensive annotated video regions paired with language queries. To eliminate the annotation costs, we make a first attempt to train a natural language video localization model in zero-shot manner. Inspired by unsupervised image captioning setup, we merely require random text corpora, unlabeled video collections, and an off-the-shelf object detector to train a model. With the unrelated and unpaired data, we propose to generate pseudo-supervision of candidate temporal regions and corresponding query sentences, and develop a simple NLVL model to train with the pseudo-supervision. Our empirical validations show that the proposed pseudo-supervised method outperforms several baseline approaches and a number of methods using stronger supervision on Charades-STA and ActivityNet-Captions.
****

Deep Halftoning With Reversible Binary Pattern
Menghan Xia, Wenbo Hu, Xueting Liu, Tien-Tsin Wong; Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), 2021, pp. 14000-14009
Existing halftoning algorithms usually drop colors and fine details when ditheri ng color images with binary dot patterns, which makes it extremely difficult to recover the original information. To dispense the recovery trouble in future, we propose a novel halftoning technique that converts a color image into binary ha lftone with full restorability to the original version. The key idea is to impli citly embed those previously dropped information into the halftone patterns. So, the halftone pattern not only serves to reproduce the image tone, maintain the blue-noise randomness, but also represents the color information and fine detail s. To this end, we exploit two collaborative convolutional neural networks (CNNs) to learn the dithering scheme, under a non-trivial self-supervision formulatio n. To tackle the flatness degradation issue of CNNs, we propose a novel noise in centive block (NIB) that can serve as a generic CNN plug-in for performance prom otion. At last, we tailor a guiding-aware training scheme that secures the conve rgence direction as regulated. We evaluate the invertible halftones in multiple aspects, which evidences the effectiveness of our method.
****

Robustness via Cross-Domain Ensembles
Teresa Yeo, O■uzhan Fatih Kar, Amir Zamir; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 12189-12199
We present a method for making neural network predictions robust to shifts from the training data distribution. The proposed method is based on making predictio ns via a diverse set of cues (called `middle domains') and ensembling them into one strong prediction. The premise of the idea is that predictions made via diff erent cues respond differently to a distribution shift, hence one should be able to merge them into one robust final prediction. We perform the merging in a str aightforward but principled manner based on the uncertainty associated with each prediction. The evaluations are performed using multiple tasks and datasets (Ta skonomy, Replica, ImageNet, CIFAR) under a wide range of adversarial and non-adv ersarial distribution shifts which demonstrate the proposed method is considerab ly more robust than its standard learning counterpart, conventional deep ensembl es, and several other baselines.
****

Topic Scene Graph Generation by Attention Distillation From Caption
Wenbin Wang, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15900-15910
If an image tells a story, the image caption is the briefest narrator. Generally , a scene graph prefers to be an omniscient "generalist", while the image captio n is more willing to be a "specialist", which outlines the gist. Lots of previou s studies have found that a scene graph is not as practical as expected unless i t can reduce the trivial contents and noises. In this respect, the image caption is a good tutor. To this end, we let the scene graph borrow the ability from th e image caption so that it can be a specialist on the basis of remaining all-aro und, resulting in the so-called Topic Scene Graph. What an image caption pays at tention to is distilled and passed to the scene graph for estimating the importa nce of partial objects, relationships, and events. Specifically, during the capt ion generation, the attention about individual objects in each time step is coll ected, pooled, and assembled to obtain the attention about relationships, which serves as weak supervision for regularizing the estimated importance scores of r elationships. In addition, as this attention distillation process provides an op portunity for combining the generation of image caption and scene graph together , we further transform the scene graph into linguistic form with rich and free-f orm expressions by sharing a single generation model with image caption. Experim ents show that attention distillation brings significant improvements in mining important relationships without strong supervision, and the topic scene graph sh ows great potential in subsequent applications.
****

FFT-OT: A Fast Algorithm for Optimal Transportation
Na Lei, Xianfeng Gu; Proceedings of the IEEE/CVF International Conference on Com puter Vision (ICCV), 2021, pp. 6280-6289

An optimal transportation map finds the most economical way to transport one probability measure to the other. It has been applied in a broad range of applications in vision, deep learning and medical images. By Brenier theory, computing the optimal transport map is equivalent to solving a Monge-Ampere equation. Due to the highly non-linear nature, the computation of optimal transportation maps in large scale is very challenging. This work proposes a simple but powerful method, the FFT-OT algorithm, to tackle this difficulty based on three key ideas. First, solving Monge-Ampere equation is converted to a fixed point problem; Second, the obliqueness property of optimal transportation maps are reformulated as Neumann boundary conditions on rectangular domains; Third, FFT is applied in each iteration to solve a Poisson equation in order to improve the efficiency. Experiments on surfaces captured from 3D scanning and reconstructed from medical imaging are conducted, and compared with other existing methods. Our experimental results show that the proposed FFT-OT algorithm is simple, general and scalable with high efficiency and accuracy.

****

Contrastive Learning for Label Efficient Semantic Segmentation

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, Ying Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10623-10633

Collecting labeled data for the task of semantic segmentation is expensive and time-consuming, as it requires dense pixel-level annotations. While recent Convolutional Neural Network (CNN) based semantic segmentation approaches have achieved impressive results by using large amounts of labeled training data, their performance drops significantly as the amount of labeled data decreases. This happens because deep CNNs trained with the de facto cross-entropy loss can easily overfit to small amounts of labeled data. To address this issue, we propose a simple and effective contrastive learning-based training strategy in which we first pretrain the network using a pixel-wise, label-based contrastive loss, and then fine-tune it using the cross-entropy loss. This approach increases intra-class compactness and inter-class separability, thereby resulting in a better pixel classifier. We demonstrate the effectiveness of the proposed training strategy using the Cityscapes and PASCAL VOC 2012 segmentation datasets. Our results show that pretraining with the proposed contrastive loss results in large performance gains (more than 20% absolute improvement in some settings) when the amount of labeled data is limited. In many settings, the proposed contrastive pretraining strategy, which does not use any additional data, is able to match or outperform the widely-used ImageNet pretraining strategy that uses more than a million additional labeled images.

****

Progressive Correspondence Pruning by Consensus Learning

Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6464-6473

Correspondence pruning aims to correctly remove false matches (outliers) from an initial set of putative correspondences. The selection is challenging since putative matches are typically extremely unbalanced, largely dominated by outliers, and the random distribution of such outliers further complicates the learning process for learning-based methods. To address this issue, we propose to progressively prune the correspondences via a local-to-global consensus learning procedure. We introduce a "pruning" block that lets us identify reliable candidates among the initial matches according to consensus scores estimated using local-to-global dynamic graphs. We then achieve progressive pruning by stacking multiple pruning blocks sequentially. Our method outperforms state-of-the-arts on robust line fitting, camera pose estimation and retrieval-based image localization benchmarks by significant margins and shows promising generalization ability to different datasets and detector/descriptor combinations.

****

BiMaL: Bijective Maximum Likelihood Approach to Domain Adaptation in Semantic Scene Segmentation

Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, Khoa Luu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8548-8557

Semantic segmentation aims to predict pixel-level labels. It has become a popular task in various computer vision applications. While fully supervised segmentation methods have achieved high accuracy on large-scale vision datasets, they are unable to generalize on a new test environment or a new domain well. In this work, we first introduce a new Un- aligned Domain Score to measure the efficiency of a learned model on a new target domain in unsupervised manner. Then, we present the new Bijective Maximum Likelihood (BiMaL) loss that is a generalized form of the Adversarial Entropy Minimization without any assumption about pixel independence. We have evaluated the proposed BiMaL on two domains. The proposed BiMaL approach consistently outperforms the SOTA methods on empirical experiments on "SYNTHIA to Cityscapes", "GTA5 to Cityscapes", and "SYNTHIA to Vistas".

****

Multiscale Vision Transformers

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, Christoph Feichtenhofer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6824-6835

We present Multiscale Vision Transformers (MViT) for video and image recognition, by connecting the seminal idea of multiscale feature hierarchies with transformer models. Multiscale Transformers have several channel-resolution scale stages. Starting from the input resolution and a small channel dimension, the stages hierarchically expand the channel capacity while reducing the spatial resolution. This creates a multiscale pyramid of features with early layers operating at high spatial resolution to model simple low-level visual information, and deeper layers at spatially coarse, but complex, high-dimensional features. We evaluate this fundamental architectural prior for modeling the dense nature of visual signals for a variety of video recognition tasks where it outperforms concurrent vision transformers that rely on large scale external pre-training and are 5-10 more costly in computation and parameters. We further remove the temporal dimension and apply our model for image classification where it outperforms prior work on vision transformers. Code is available at: https://github.com/facebookresearch/SlowFast.

****

Robust Small Object Detection on the Water Surface Through Fusion of Camera and Millimeter Wave Radar

Yuwei Cheng, Hu Xu, Yimin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15263-15272

In recent years, unmanned surface vehicles (USVs) have been experiencing growth in various applications. With the expansion of USVs' application scenes from the typical marine areas to inland waters, new challenges arise for the object detection task, which is an essential part of the perception system of USVs. In our work, we focus on a relatively unexplored task for USVs in inland waters: small object detection on water surfaces, which is of vital importance for safe autonomous navigation and USVs' certain missions such as floating waste cleaning. Considering the limitations of vision-based object detection, we propose a novel vision-radar fusion based method for robust small object detection on water surfaces. By using a novel representation format of millimeter wave radar point clouds and applying a deep-level multi-scale fusion of RGB images and radar data, the proposed method can efficiently utilize the characteristics of radar data and improve the accuracy and robustness for small object detection on water surfaces. We test the method on the real-world floating bottle dataset that we collected and released. The result shows that, our method improves the average detection accuracy significantly compared to the vision-based methods and achieves state-of-the-art performance. Besides, the proposed method performs robustly when single sensor degrades.

****

Just a Few Points Are All You Need for Multi-View Stereo: A Novel Semi-Supervised Learning Method for Multi-View Stereo

Taekyung Kim, Jaehoon Choi, Seokeon Choi, Dongki Jung, Changick Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6178-6186

While learning-based multi-view stereo (MVS) methods have recently shown successful performances in quality and efficiency, limited MVS data hampers generalization to unseen environments. A simple solution is to generate various large-scale MVS datasets, but generating dense ground truth for 3D structure requires a huge amount of time and resources. On the other hand, if the reliance on dense ground truth is relaxed, MVS systems will generalize more smoothly to new environments. To this end, we first introduce a novel semi-supervised multi-view stereo framework called a Sparse Ground truth-based MVS Network (SGT-MVSNet) that can reliably reconstruct the 3D structures even with a few ground truth 3D points. Our strategy is to divide the accurate and erroneous regions and individually conquer them based on our observation that a probability map can separate these regions. We propose a self-supervision loss called the 3D Point Consistency Loss to enhance the 3D reconstruction performance, which forces the 3D points back-projected from the corresponding pixels by the predicted depth values to meet at the same 3D coordinates. Finally, we propagate these improved depth predictions toward edges and occlusions by the Coarse-to-fine Reliable Depth Propagation module. We generate the spare ground truth of the DTU dataset for evaluation and extensive experiments verify that our SGT-MVSNet outperforms the state-of-the-art MVS methods on the sparse ground truth setting. Moreover, our method shows comparable reconstruction results to the supervised MVS methods though we only used tens and hundreds of ground truth 3D points.

****

Multispectral Illumination Estimation Using Deep Unrolling Network

Yuqi Li, Qiang Fu, Wolfgang Heidrich; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2672-2681

This paper examines the problem of illumination spectra estimation in multispectral images. We cast the problem into a constrained matrix factorization problem and present a method for both single-global and multiple illumination estimation in which a deep unrolling network is constructed from the alternating direction method of multipliers(ADMM) optimization for solving the matrix factorization problem. To alleviate the lack of multispectral training data, we build a large multispectral reflectance image dataset for generating synthesized data and use them for training and evaluating our model. The results of simulations and real experiments demonstrate that the proposed method is able to outperform state-of-the-art spectral illumination estimation methods, and that it generalizes well to a wide variety of scenes and spectra.

****

GroupFormer: Group Activity Recognition With Clustered Spatial-Temporal Transformer

Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, Shuai Yi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13668-13677

Group activity recognition is a crucial yet challenging problem, whose core lies in fully exploring spatial-temporal interactions among individuals and generating reasonable group representations. However, previous methods either model spatial and temporal information separately, or directly aggregate individual features to form group features. To address these issues, we propose a novel group activity recognition network termed GroupFormer. It captures spatial-temporal contextual information jointly to augment the individual and group representations effectively with a clustered spatial-temporal transformer. Specifically, our GroupFormer has three appealing advantages: (1) A tailor-modified Transformer, Clustered Spatial-Temporal Transformer, is proposed to enhance the individual and group representation. (2) It models the spatial and temporal dependencies integrally and utilizes decoders to build the bridge between the spatial and temporal information. (3) A clustered attention mechanism is utilized to dynamically divide individuals into multiple clusters for better learning activity-aware semantic representations. Moreover, experimental results show that the proposed framework o

utperforms state-of-the-art methods on the Volleyball dataset and Collective Act
ivity dataset.
****

BAPA-Net: Boundary Adaptation and Prototype Alignment for Cross-Domain Semantic Segmentation
Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, Lixin Duan; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8801-8811
Existing cross-domain semantic segmentation methods usually focus on the overall
 segmentation results of whole objects but neglect the importance of object boun
daries. In this work, we find that the segmentation performance can be considera
bly boosted if we treat object boundaries properly. For that, we propose a novel
 method called BAPA-Net, which is based on a convolutional neural network via Bo
undary Adaptation and Prototype Alignment, under the unsupervised domain adaptat
ion setting. Specifically, we first construct additional images by pasting objec
ts from source images to target images, and we develop a so-called boundary adap
tation module to weigh each pixel based on its distance to the nearest boundary
pixel of those pasted source objects. Moreover, we pro- pose another prototype a
lignment module to reduce the domain mismatch by minimizing distances between th
e class prototypes of the source and target domains, where boundaries are remove
d to avoid domain confusion during prototype calculation. By integrating the bou
ndary adaptation and prototype alignment, we are able to train a discriminative
and domain-invariant model for cross-domain semantic segmentation. We conduct ex
tensive experiments on the benchmark datasets of urban scenes (i.e., GTA5->Citys
capes and SYNTHIA->Cityscapes). And the promising results clearly show the effec
tiveness of our BAPA-Net method over existing state-of-the-art for cross-domain
semantic segmentation. Our implementation is available at https://github.com/man
manjun/BAPA-Net.
****

TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild
Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehs
an Adeli, Hamid Rezatofighi; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 13390-13400
Joint forecasting of human trajectory and pose dynamics is a fundamental buildin
g block of various applications ranging from robotics and autonomous driving to
surveillance systems. Predicting body dynamics requires capturing subtle informa
tion embedded in the humans' interactions with each other and with the objects p
resent in the scene. In this paper, we propose a novel TRajectory and POse Dynam
ics (nicknamed TRiPOD) method based on graph attentional networks to model the h
uman-human and human-object interactions both in the input space and the output
space (decoded future output). The model is supplemented by a message passing in
terface over the graphs to fuse these different levels of interactions efficient
ly. Furthermore, to incorporate a real-world challenge, we propound to learn an
indicator representing whether an estimated body joint is visible/invisible at e
ach frame, e.g. due to occlusion or being outside the sensor field of view. Fina
lly, we introduce a new benchmark for this joint task based on two challenging d
atasets (PoseTrack and 3DPW) and propose evaluation metrics to measure the effec
tiveness of predictions in the global space, even when there are invisible cases
 of joints. Our evaluation shows that TRiPOD outperforms all prior work and stat
e-of-the-art specifically designed for each of the trajectory and pose forecasti
ng tasks.
****

Conditional DETR for Fast Training Convergence
Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun
, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2021, pp. 3651-3660
The recently-developed DETR approach applies the transformer encoder and decoder
 architecture to object detection and achieves promising performance. In this pa
per, we handle the critical issue, slow training convergence, and present a cond
itional cross-attention mechanism for fast DETR training. Our approach is motiva
ted by that the cross-attention in DETR relies highly on the content embeddings

for localizing the four extremities and predicting the box, which increases the need for high-quality content embeddings and thus the training difficulty. Our approach, named conditional DETR, learns a conditional spatial query from the decoder embedding for decoder multi-head cross-attention. The benefit is that through the conditional spatial query, each cross-attention head is able to attend to a band containing a distinct region, e.g., one object extremity or a region inside the object box. This narrows down the spatial range for localizing the distinct regions for object classification and box regression, thus relaxing the dependence on the content embeddings and easing the training. Empirical results show that conditional DETR converges 6.7x faster for the backbones R50 and R101 and 10x faster for stronger backbones DC5-R50 and DC5-R101. Code is available at https://github.com/Atten4Vis/ConditionalDETR.

****

Distilling Global and Local Logits With Densely Connected Relations

Youmin Kim, Jinbae Park, YounHo Jang, Muhammad Ali, Tae-Hyun Oh, Sung-Ho Bae; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6290-6300

In prevalent knowledge distillation, logits in most image recognition models are computed by global average pooling, then used to learn to encode the high-level and task-relevant knowledge. In this work, we solve the limitation of this global logit transfer in this distillation context. We point out that it prevents the transfer of informative spatial information, which provides localized knowledge as well as rich relational information across contexts of an input scene. To exploit the rich spatial information, we propose a simple yet effective logit distillation approach. We add a local spatial pooling layer branch to the penultimate layer, thereby our method extends the standard logit distillation and enables learning of both finely-localized knowledge and holistic representation. Our proposed method shows favorable accuracy improvement against the state-of-the-art methods on several image classification datasets. We show that our distilled students trained on the image classification task can be successfully leveraged for object detection and semantic segmentation tasks; this result demonstrates our method's high transferability.

****

A Hierarchical Transformation-Discriminating Generative Model for Few Shot Anomaly Detection

Shelly Sheynin, Sagie Benaim, Lior Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8495-8504

Anomaly detection, the task of identifying unusual samples in data, often relies on a large set of training samples. In this work, we consider the setting of few-shot anomaly detection in images, where only a few images are given at training. We devise a hierarchical generative model that captures the multi-scale patch distribution of each training image. We further enhance the representation of our model by using image transformations and optimize scale-specific patch-discriminators to distinguish between real and fake patches of the image, as well as between different transformations applied to those patches. The anomaly score is obtained by aggregating the patch-based votes of the correct transformation across scales and image regions. We demonstrate the superiority of our method on both the one-shot and few-shot settings, on the datasets of Paris, CIFAR10, MNIST and FashionMNIST as well as in the setting of defect detection on MVTec. In all cases, our method outperforms the recent baseline methods.

****

MVTN: Multi-View Transformation Network for 3D Shape Recognition

Abdullah Hamdi, Silvio Giancola, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1-11

Multi-view projection methods have demonstrated their ability to reach state-of-the-art performance on 3D shape recognition. Those methods learn different ways to aggregate information from multiple views. However, the camera view-points for those views tend to be heuristically set and fixed for all shapes. To circumvent the lack of dynamism of current multi-view methods, we propose to learn those view-points. In particular, we introduce the Multi-View Transformation Network

(MVTN) that regresses optimal view-points for 3D shape recognition, building upo
n advances in differentiable rendering. As a result, MVTN can be trained end-to-
end along with any multi-view network for 3D shape classification. We integrate
MVTN in a novel adaptive multi-view pipeline that can render either 3D meshes or
 point clouds. MVTN exhibits clear performance gains in the tasks of 3D shape cl
assification and 3D shape retrieval without the need for extra training supervis
ion. In these tasks, MVTN achieves state-of-the-art performance on ModelNet40, S
hapeNet Core55, and the most recent and realistic ScanObjectNN dataset (up to 6
% improvement). Interestingly, we also show that MVTN can provide network robust
ness against rotation and occlusion in the 3D domain.
****

GNeRF: GAN-Based Neural Radiance Field Without Posed Camera
Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, Jingyi Y
u; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2021, pp. 6351-6361
We introduce GNeRF, a framework to marry Generative Adversarial Networks (GAN) w
ith Neural Radiance Field (NeRF) reconstruction for the complex scenarios with u
nknown and even randomly initialized camera poses. Recent NeRF-based advances ha
ve gained popularity for remarkable realistic novel view synthesis. However, mos
t of them heavily rely on accurate camera poses estimation, while few recent met
hods can only optimize the unknown camera poses in roughly forward-facing scenes
 with relatively short camera trajectories and require rough camera poses initia
lization. Differently, our GNeRF only utilizes randomly initialized poses for co
mplex outside-in scenarios. We propose a novel two-phases end-to-end framework.
The first phase takes the use of GANs into the new realm for optimizing coarse c
amera poses and radiance fields jointly, while the second phase refines them wit
h additional photometric loss. We overcome local minima using a hybrid and itera
tive optimization scheme. Extensive experiments on a variety of synthetic and na
tural scenes demonstrate the effectiveness of GNeRF. More impressively, our appr
oach outperforms the baselines favorably in those scenes with repeated patterns
or even low textures that are regarded as extremely challenging before.
****

ODAM: Object Detection, Association, and Mapping Using Posed RGB Video
Kejie Li, Daniel DeTone, Yu Fan (Steven) Chen, Minh Vo, Ian Reid, Hamid Rezatofi
ghi, Chris Sweeney, Julian Straub, Richard Newcombe; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 5998-6008
Localizing objects and estimating their extent in 3D is an important step toward
s high-level 3D scene understanding, which has many applications in Augmented Re
ality and Robotics. We present ODAM, a system for 3D Object Detection, Associati
on, and Mapping using posed RGB videos. The proposed system relies on a deep-lea
rning-based front-end to detect 3D objects from a given RGB frame and associate
them to a global object-based map using a graph neural network (GNN). Based on t
hese frame-to-model associations, our back-end optimizes object bounding volumes
, represented as super-quadrics, under multi-view geometry constraints and the o
bject scale prior. We validate the proposed system on ScanNet where we show a si
gnificant improvement over existing RGB-only methods.
****

Learning Specialized Activation Functions With the Piecewise Linear Unit
Yucong Zhou, Zezhou Zhu, Zhao Zhong; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2021, pp. 12095-12104
The choice of activation functions is crucial for modern deep neural networks. P
opular hand-designed activation functions like Rectified Linear Unit(ReLU) and i
ts variants show promising performance in various tasks and models. Swish, the a
utomatically discovered activation function, outperforms ReLU on many challengin
g datasets. However, it has two main drawbacks. First, the tree-based search spa
ce is highly discrete and restricted, making it difficult to searching. Second,
the sample-based searching method is inefficient, making it infeasible to find s
pecialized activation functions for each dataset or neural architecture. To tack
le these drawbacks, we propose a new activation function called Piecewise Linear
 Unit(PWLU), which incorporates a carefully designed formulation and learning me

thod. It can learn specialized activation functions and achieves SOTA performanc e on large-scale datasets like ImageNet and COCO. For example, on ImageNet class ification dataset, PWLU improves 0.9%/0.53%/1.0%/1.7%/1.0% top-1 accuracy over S wish for ResNet-18/ResNet-50/MobileNet-V2/MobileNet-V3/EfficientNet-B0. PWLU is also easy to implement and efficient at inference, which can be widely applied i n real-world applications.
****

Viewpoint Invariant Dense Matching for Visual Geolocalization
Gabriele Berton, Carlo Masone, Valerio Paolicelli, Barbara Caputo; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 121 69-12178
In this paper we propose a novel method for image matching based on dense local features and tailored for visual geolocalization. Dense local features matching is robust against changes in illumination and occlusions, but not against viewpo int shifts which are a fundamental aspect of geolocalization. Our method, called GeoWarp, directly embeds invariance to viewpoint shifts in the process of extra cting dense features. This is achieved via a trainable module which learns from the data an invariance that is meaningful for the task of recognizing places. We also devise a new self-supervised loss and two new weakly supervised losses to train this module using only unlabeled data and weak labels. GeoWarp is implemen ted efficiently as a re-ranking method that can be easily embedded into pre-exis ting visual geolocalization pipelines. Experimental validation on standard geolo calization benchmarks demonstrates that GeoWarp boosts the accuracy of state-of- the-art retrieval architectures. The code and trained models will be released up on acceptance of this paper.
****

Dual Contrastive Loss and Attention for GANs
Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S. Davis , Mario Fritz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6731-6742
Generative Adversarial Networks (GANs) produce impressive results on uncondition al image generation when powered with large-scale image datasets. Yet generated images are still easy to spot especially on datasets with high variance (e.g. be droom, church). In this paper, we propose various improvements to further push t he boundaries in image generation. Specifically, we propose a novel dual contras tive loss and show that, with this loss, discriminator learns more generalized a nd distinguishable representations to incentivize generation. In addition, we re visit attention and extensively experiment with different attention blocks in th e generator. We find attention to be still an important module for successful im age generation even though it was not used in the recent state-of-the-art models . Lastly, we study different attention architectures in the discriminator, and p ropose a reference attention mechanism. By combining the strengths of these reme dies, we improve the compelling state-of-the-art Frechet Inception Distance (FID) by at least 17.5% on several benchmark datasets. We obtain even more significa nt improvements on compositional synthetic scenes (up to 47.5% in FID).
****

Video Autoencoder: Self-Supervised Disentanglement of Static 3D Structure and Mo tion
Zihang Lai, Sifei Liu, Alexei A. Efros, Xiaolong Wang; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 9730-9740
We present Video Autoencoder for learning disentangled representations of 3D str ucture and camera pose from videos in a self-supervised manner. Relying on tempo ral continuity in videos, our work assumes that the 3D scene structure in nearby video frames remains static. Given a sequence of video frames as input, the Vid eo Autoencoder extracts a disentangled representation of the scene including: (i) a temporally-consistent deep voxel feature to represent the 3D structure and (ii) a 3D trajectory of camera poses for each frame. These two representations wi ll then be re-entangled for rendering the input video frames. Video Autoencoder can be trained directly using a pixel reconstruction loss, without any ground tr uth 3D or camera pose annotations. The disentangled representation can be applie

d to a range of tasks, including novel view synthesis, camera pose estimation, a
nd video generation by motion following. We evaluate our method on several large
-scale natural video datasets, and show generalization results on out-of-domain
images.
****

Adaptive Convolutions With Per-Pixel Dynamic Filter Atom
Ze Wang, Zichen Miao, Jun Hu, Qiang Qiu; Proceedings of the IEEE/CVF Internation
al Conference on Computer Vision (ICCV), 2021, pp. 12302-12311
Applying feature dependent network weights have been proved to be effective in m
any fields. However, in practice, restricted by the enormous size of model param
eters and memory footprints, scalable and versatile dynamic convolutions with pe
r-pixel adapted filters are yet to be fully explored. In this paper, we address
this challenge by decomposing filters, adapted to each spatial position, over dy
namic filter atoms generated by a light-weight network from local features. Adap
tive receptive fields can be supported by further representing each filter atom
over sets of pre-fixed multi-scale bases. As plug-and-play replacements to convo
lutional layers, the introduced adaptive convolutions with per-pixel dynamic ato
ms enable explicit modeling of intra-image variance, while avoiding heavy comput
ation, parameters, and memory cost. Our method preserves the appealing propertie
s of conventional convolutions as being translation-equivariant and parametrical
ly efficient. We present experiments to show that, the proposed method delivers
comparable or even better performance across tasks, and are particularly effecti
ve on handling tasks with significant intra-image variance.
****

Video Pose Distillation for Few-Shot, Fine-Grained Sports Action Recognition
James Hong, Matthew Fisher, Michaël Gharbi, Kayvon Fatahalian; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9254-92
63
Human pose is a useful feature for fine-grained sports action understanding. How
ever, pose estimators are often unreliable when run on sports video due to domai
n shift and factors such as motion blur and occlusions. This leads to poor accur
acy when downstream tasks, such as action recognition, depend on pose. End-to-en
d learning circumvents pose, but requires more labels to generalize. We introduc
e Video Pose Distillation (VPD), a weakly-supervised technique to learn features
 for new video domains, such as individual sports that challenge pose estimation
. Under VPD, a student network learns to extract robust pose features from RGB f
rames in the sports video, such that, whenever pose is considered reliable, the
features match the output of a pretrained teacher pose detector. Our strategy re
tains the best of both pose and end-to-end worlds, exploiting the rich visual pa
tterns in raw video frames, while learning features that agree with the athletes
' pose and motion in the target video domain to avoid over-fitting to patterns u
nrelated to athletes' motion. VPD features improve performance on few-shot, fine
-grained action recognition, retrieval, and detection tasks in four real-world s
ports video datasets, without requiring additional ground-truth pose annotations
.
****

Else-Net: Elastic Semantic Network for Continual Action Recognition From Skeleto
n Data
Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, Jun Liu; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
, pp. 13434-13443
We address continual action recognition from skeleton sequence, which aims to le
arn a recognition model over time from a continuous stream of skeleton data. Thi
s task is very important in changing environment. Due to catastrophic forgetting
 problems of deep neural networks and large discrepancies between the previously
 learned and current new human actions from different categories, the neural net
works may "forget" old actions, when learning new actions. This makes online con
tinual action recognition a challenging task. We observe that although different
 human actions may vary to a large extent as a whole, their local body parts cou
ld share similar features. Therefore, we propose an Elastic Semantic Network (El

se-Net) to learn new actions by decomposing human bodies into several semantic b ody parts. For each body part, the proposed Else-Net constructs a semantic pathw ay using several elastic cells learned with old actions, or explores new cells t o store new knowledge.

****

Low-Shot Validation: Active Importance Sampling for Estimating Classifier Perfor mance on Rare Categories

Fait Poms, Vishnu Sarukkai, Ravi Teja Mullapudi, Nimit S. Sohoni, William R. Mar k, Deva Ramanan, Kayvon Fatahalian; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2021, pp. 10705-10714

For machine learning models trained with limited labeled training data, validati on stands to become the main bottleneck to reducing overall annotation costs. We propose a statistical validation algorithm that accurately estimates the F-scor e of binary classifiers for rare categories, where finding relevant examples to evaluate on is particularly challenging. Our key insight is that simultaneous ca libration and importance sampling enables accurate estimates even in the low-sam ple regime (<300 samples). Critically, we also derive an accurate single-trial e stimator of the variance of our method and demonstrate that this estimator is em pirically accurate at low sample counts, enabling a practitioner to know how wel l they can trust a given low-sample estimate. When validating state-of-the-art s emi-supervised models on ImageNet and iNaturalist2017, our method achieves the s ame estimates of model performance with up to 10x fewer labels than competing ap proaches. In particular, we can estimate model F1 scores with a variance of 0.00 5 using as few as 100 labels.

****

Deep Matching Prior: Test-Time Optimization for Dense Correspondence

Sunghwan Hong, Seungryong Kim; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 9907-9917

Conventional techniques to establish dense correspondences across visually or se mantically similar images focused on designing a task-specific matching prior, w hich is difficult to model in general. To overcome this, recent learning-based m ethods have attempted to learn a good matching prior within a model itself on la rge training data. The performance improvement was apparent, but the need for su fficient training data and intensive learning hinders their applicability. Moreo ver, using the fixed model at test time does not account for the fact that a pai r of images may require their own prior, thus providing limited performance and poor generalization to unseen images. In this paper, we show that an image pair-specific prior can be captured by solely optimizing the untrained matching netwo rks on an input pair of images. Tailored for such test-time optimization for den se correspondence, we present a residual matching network and a confidence-aware contrastive loss to guarantee a meaningful convergence. Experiments demonstrate that our framework, dubbed Deep Matching Prior (DMP), is competitive, or even o utperforms, against the latest learning-based methods on several benchmarks for geometric matching and semantic matching, even though it requires neither large training data nor intensive learning. With the networks pre-trained, DMP attains state-of-the-art performance on all benchmarks.

****

DualPoseNet: Category-Level 6D Object Pose and Size Estimation Using Dual Pose N etwork With Refined Learning of Pose Consistency

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, Yuanqing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3 560-3569

Category-level 6D object pose and size estimation is to predict full pose config urations of rotation, translation, and size for object instances observed in sin gle, arbitrary views of cluttered scenes. In this paper, we propose a new method of Dual Pose Network with refined learning of pose consistency for this task, s hortened as DualPoseNet. DualPoseNet stacks two parallel pose decoders on top of a shared pose encoder, where the implicit decoder predicts object poses with a working mechanism different from that of the explicit one; they thus impose comp lementary supervision on the training of pose encoder. We construct the encoder

based on spherical convolutions, and design a module of Spherical Fusion wherein for a better embedding of pose-sensitive features from the appearance and shape observations. Given no testing CAD models, it is the novel introduction of the implicit decoder that enables the refined pose prediction during testing, by enforcing the predicted pose consistency between the two decoders using a self-adaptive loss term. Thorough experiments on benchmarks of both category- and instance-level object pose datasets confirm efficacy of our designs. DualPoseNet outperforms existing methods with a large margin in the regime of high precision. Our code is released publicly at https://github.com/Gorilla-Lab-SCUT/DualPoseNet.
****

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding
Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, Nicolas Carion; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1780-1790
Multi-modal reasoning systems rely on a pre-trained object detector to extract regions of interest from the image. However, this crucial module is typically used as a black box, trained independently of the downstream task and on a fixed vocabulary of objects and attributes. This makes it challenging for such systems to capture the long tail of visual concepts expressed in free form text. In this paper we propose MDETR, an end-to-end modulated detector that detects objects in an image conditioned on a raw text query, like a caption or a question. We use a transformer-based architecture to reason jointly over text and image by fusing the two modalities at an early stage of the model. We pre-train the network on 1.3M text-image pairs, mined from pre-existing multi-modal datasets having explicit alignment between phrases in text and objects in the image. We then fine-tune on several downstream tasks such as phrase grounding, referring expression comprehension and segmentation, achieving state-of-the-art results on popular benchmarks. We also investigate the utility of our model as an object detector on a given label set when fine-tuned in a few-shot setting. We show that our pre-training approach provides a way to handle the long tail of object categories which have very few labelled instances. Our approach can be easily extended for visual question answering, achieving competitive performance on GQA and CLEVR. The code and models are available at https://github.com/ashkamath/mdetr.
****

Calibrated and Partially Calibrated Semi-Generalized Homographies
Snehal Bhayani, Torsten Sattler, Daniel Barath, Patrik Beliansky, Janne Heikkilä, Zuzana Kukelova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5936-5945
In this paper, we propose the first minimal solutions for estimating the semi-generalized homography given a perspective and a generalized camera. The proposed solvers use five 2D-2D image point correspondences induced by a scene plane. One group of solvers assumes the perspective camera to be fully calibrated, while the other estimates the unknown focal length together with the absolute pose parameters. This setup is particularly important in structure-from-motion and visual localization pipelines, where a new camera is localized in each step with respect to a set of known cameras and 2D-3D correspondences might not be available. Thanks to a clever parametrization and the elimination ideal method, our solvers only need to solve a univariate polynomial of degree five or three, respectively a system of polynomial equations in two variables. All proposed solvers are stable and efficient as demonstrated by a number of synthetic and real-world experiments.
****

End-to-End Video Instance Segmentation via Spatial-Temporal Graph Neural Networks
Tao Wang, Ning Xu, Kean Chen, Weiyao Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10797-10806
Video instance segmentation is a challenging task that extends image instance segmentation to the video domain. Existing methods either rely only on single-frame information for the detection and segmentation subproblems or handle tracking as a separate post-processing step, which limit their capability to fully levera

ge and share useful spatial-temporal information for all the subproblems. In thi s paper, we propose a novel graph-neural-network (GNN) based method to handle th e aforementioned limitation. Specifically, graph nodes representing instance fea tures are used for detection and segmentation while graph edges representing ins tance relations are used for tracking. Both inter and intra-frame information is effectively propagated and shared via graph updates and all the subproblems (i. e. detection, segmentation and tracking) are jointly optimized in an unified fra mework. The performance of our method shows great improvement on the YoutubeVIS validation dataset compared to existing methods and achieves 36.5% AP with a Res Net-50 backbone, operating at 22 FPS.
****

The Surprising Impact of Mask-Head Architecture on Novel Class Segmentation
Vighnesh Birodkar, Zhichao Lu, Siyang Li, Vivek Rathod, Jonathan Huang; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 7015-7025
Instance segmentation models today are very accurate when trained on large annot ated datasets, but collecting mask annotations at scale is prohibitively expensi ve. We address the partially supervised instance segmentation problem in which o ne can train on (significantly cheaper) bounding boxes for all categories but us e masks only for a subset of categories. In this work, we focus on a popular fam ily of models which apply differentiable cropping to a feature map and predict a mask based on the resulting crop. Under this family, we study Mask R-CNN and di scover that instead of its default strategy of training the mask-head with a com bination of proposals and groundtruth boxes, training the mask-head with only gr oundtruth boxes dramatically improves its performance on novel classes. This tra ining strategy also allows us to take advantage of alternative mask-head archite ctures, which we exploit by replacing the typical mask-head of 2-4 layers with s ignificantly deeper off-the-shelf architectures (e.g. ResNet, Hourglass models). While many of these architectures perform similarly when trained in fully super vised mode, our main finding is that they can generalize to novel classes in dra matically different ways. We call this ability of mask-heads to generalize to un seen classes the strong mask generalization effect and show that without any spe cialty modules or losses, we can achieve state-of-the-art results in the partial ly supervised COCO instance segmentation benchmark. Finally, we demonstrate that our effect is general, holding across underlying detection methodologies (inclu ding anchor-based, anchor-free or no detector at all) and across different backb one networks. Code and pre-trained models are available at https://git.io/deepma c.
****

The Spatio-Temporal Poisson Point Process: A Simple Model for the Alignment of E vent Camera Data
Cheng Gu, Erik Learned-Miller, Daniel Sheldon, Guillermo Gallego, Pia Bideau; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 13495-13504
Event cameras, inspired by biological vision systems, provide a natural and data efficient representation of visual information. Visual information is acquired in the form of events that are triggered by local brightness changes. However, b ecause most brightness changes are triggered by relative motion of the camera an d the scene, the events recorded at a single sensor location seldom correspond t o the same world point. To extract meaningful information from event cameras, it is helpful to register events that were triggered by the same underlying world point. In this work we propose a new model of event data that captures its natur al spatio-temporal structure. We start by developing a model for aligned event d ata. That is, we develop a model for the data as though it has been perfectly re gistered already. In particular, we model the aligned data as a spatio-temporal Poisson point process. Based on this model, we develop a maximum likelihood appr oach to registering events that are not yet aligned. That is, we find transforma tions of the observed events that make them as likely as possible under our mode l. In particular we extract the camera rotation that leads to the best event ali gnment. We show new state of the art accuracy for rotational velocity estimation

on the DAVIS 240C dataset [??]. In addition, our method is also faster and has lower computational complexity than several competing methods.
****

Learning Self-Similarity in Space and Time As Generalized Motion for Video Action Recognition
Heeseung Kwon, Manjin Kim, Suha Kwak, Minsu Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13065-13075
Spatio-temporal convolution often fails to learn motion dynamics in videos and thus an effective motion representation is required for video understanding in the wild. In this paper, we propose a rich and robust motion representation based on spatio-temporal self-similarity (STSS). Given a sequence of frames, STSS represents each local region as similarities to its neighbors in space and time. By converting appearance features into relational values, it enables the learner to better recognize structural patterns in space and time. We leverage the whole volume of STSS and let our model learn to extract an effective motion representation from it. The proposed neural block, dubbed SELFY, can be easily inserted into neural architectures and trained end-to-end without additional supervision. With a sufficient volume of the neighborhood in space and time, it effectively captures long-term interaction and fast motion in the video, leading to robust action recognition. Our experimental analysis demonstrates its superiority over previous methods for motion modeling as well as its complementarity to spatio-temporal features from direct convolution. On the standard action recognition benchmarks, SomethingSomething-V1 & V2, Diving-48, and FineGym, the proposed method achieves the state-of-the-art results.
****

Collaborative Optimization and Aggregation for Decentralized Domain Generalization and Adaptation
Guile Wu, Shaogang Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6484-6493
Contemporary domain generalization (DG) and multi-source unsupervised domain adaptation (UDA) methods mostly collect data from multiple domains together for joint optimization. However, this centralized training paradigm poses a threat to data privacy and is not applicable when data are non-shared across domains. In this work, we propose a new approach called Collaborative Optimization and Aggregation (COPA), which aims at optimizing a generalized target model for decentralized DG and UDA, where data from different domains are non-shared and private. Our base model consists of a domain-invariant feature extractor and an ensemble of domain-specific classifiers. In an iterative learning process, we optimize a local model for each domain, and then centrally aggregate local feature extractors and assemble domain-specific classifiers to construct a generalized global model, without sharing data from different domains. To improve generalization of feature extractors, we employ hybrid batch-instance normalization and collaboration of frozen classifiers. For better decentralized UDA, we further introduce a prediction agreement mechanism to overcome local disparities towards central model aggregation. Extensive experiments on five DG and UDA benchmark datasets show that COPA is capable of achieving comparable performance against the state-of-the-art DG and UDA methods without the need for centralized data collection in model training.
****

CR-Fill: Generative Image Inpainting With Auxiliary Contextual Reconstruction
Yu Zeng, Zhe Lin, Huchuan Lu, Vishal M. Patel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14164-14173
Recent deep generative inpainting methods use attention layers to allow the generator to explicitly borrow feature patches from the known region to complete a missing region. Due to the lack of supervision signals for the correspondence between missing regions and known regions, it may fail to find proper reference features, which often leads to artifacts in the results. Also, it computes pair-wise similarity across the entire feature map during inference bringing a significant computational overhead. To address this issue, we propose to teach such patch-borrowing behavior to an attention-free generator by joint training of an auxil

iary contextual reconstruction task, which encourages the generated output to be plausible even when reconstructed by surrounding regions. The auxiliary branch can be seen as a learnable loss function, i.e. named as contextual reconstruction (CR) loss, where query-reference feature similarity and reference-based reconstructor are jointly optimized with the inpainting generator. The auxiliary branch (i.e. CR loss) is required only during training, and only the inpainting generator is required during the inference. Experimental results demonstrate that the proposed inpainting model compares favourably against the state-of-the-art in terms of quantitative and visual performance. Code is available at https://github.com/zengxianyu/crfill.

****

LookOut: Diverse Multi-Future Prediction and Planning for Self-Driving
Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16107-16116
In this paper, we present LookOut, a novel autonomy system that perceives the environment, predicts a diverse set of futures of how the scene might unroll and estimates the trajectory of the SDV by optimizing a set of contingency plans over these future realizations. In particular, we learn a diverse joint distribution over multi-agent future trajectories in a traffic scene that covers a wide range of future modes with high sample efficiency while leveraging the expressive power of generative models. Unlike previous work in diverse motion forecasting, our diversity objective explicitly rewards sampling future scenarios that require distinct reactions from the self-driving vehicle for improved safety. Our contingency planner then finds comfortable and non-conservative trajectories that ensure safe reactions to a wide range of future scenarios. Through extensive evaluations, we show that our model demonstrates significantly more diverse and sample-efficient motion forecasting in a large-scale self-driving dataset as well as safer and less conservative motion plans in long-term closed-loop simulations when compared to current state-of-the-art models.

****

Causal Attention for Unbiased Visual Recognition
Tan Wang, Chang Zhou, Qianru Sun, Hanwang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3091-3100
Attention module does not always help deep models learn causal features that are robust in any confounding context, e.g., a foreground object feature is invariant to different backgrounds. This is because the confounders trick the attention to capture spurious correlations that benefit the prediction when the training and testing data are IID (identical & independent distribution); while harm the prediction when the data are OOD (out-of-distribution). The sole fundamental solution to learn causal attention is by causal intervention, which requires additional annotations of the confounders, e.g., a "dog" model is learned within "grass+dog" and "road+dog" respectively, so the "grass" and "road" contexts will no longer confound the "dog" recognition. However, such annotation is not only prohibitively expensive, but also inherently problematic, as the confounders are elusive in nature. In this paper, we propose a causal attention module (CaaM) that self-annotates the confounders in unsupervised fashion. In particular, multiple CaaMs can be stacked and integrated in conventional attention CNN and self-attention Vision Transformer. In OOD settings, deep models with CaaM outperform those without it significantly; even in IID settings, the attention localization is also improved by CaaM, showing a great potential in applications that require robust visual saliency. Codes are available at https://github.com/Wangt-CN/CaaM.

****

EC-DARTS: Inducing Equalized and Consistent Optimization Into DARTS
Qinqin Zhou, Xiawu Zheng, Liujuan Cao, Bineng Zhong, Teng Xi, Gang Zhang, Errui Ding, Mingliang Xu, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11986-11995
Based on the relaxed search space, differential architecture search (DARTS) is efficient in searching for a high-performance architecture. However, the unbalanced competition among operations that have different trainable parameters causes

the model collapse. Besides, the inconsistent structures in the search and retra
ining stages causes cross-stage evaluation to be unstable. In this paper, we cal
l these issues as an operation gap and a structure gap in DARTS. To shrink these
 gaps, we propose to induce equalized and consistent optimization in differentia
ble architecture search (EC-DARTS). EC-DARTS decouples different operations base
d on their categories to optimize the operation weights so that the operation ga
p between them is shrinked. Besides, we introduce an induced structural transiti
on to bridge the structure gap between the model structures in the search and re
training stages. Extensive experiments on CIFAR10 and ImageNet demonstrate the e
ffectiveness of our method. Specifically, on CIFAR10, we achieve a test error of
 2.39%, while only 0.3 GPU days on NVIDIA TITAN V. On ImageNet, our method achie
ves a top-1 error of 23.6% under the mobile setting.
****

Detecting Persuasive Atypicality by Modeling Contextual Compatibility
Meiqi Guo, Rebecca Hwa, Adriana Kovashka; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2021, pp. 972-982
We propose a new approach to detect atypicality in persuasive imagery. Unlike at
ypicality which has been studied in prior work, persuasive atypicality has a par
ticular purpose to convey meaning, and relies on understanding the common-sense
spatial relations of objects. We propose a self-supervised attention-based techn
ique which captures contextual compatibility, and models spatial relations in a
precise manner. We further experiment with capturing common sense through the se
mantics of co-occurring object classes. We verify our approach on a dataset of a
typicality in visual advertisements, as well as a second dataset capturing atypi
cality that has no persuasive intent.
****

Warp-Refine Propagation: Semi-Supervised Auto-Labeling via Cycle-Consistency
Aditya Ganeshan, Alexis Vallet, Yasunori Kudo, Shin-ichi Maeda, Tommi Kerola, Ra
res Ambrus, Dennis Park, Adrien Gaidon; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2021, pp. 15499-15509
Deep learning models for semantic segmentation rely on expensive, large-scale, m
anually annotated datasets. Labelling is a tedious process that can take hours p
er image. Automatically annotating video sequences by propagating sparsely label
ed frames through time is a more scalable alternative. In this work, we propose
a novel label propagation method, termed Warp-Refine Propagation, that combines
semantic cues with geometric cues to efficiently auto-label videos. Our method l
earns to refine geometrically-warped labels and infuse them with learned semanti
c priors in a semi-supervised setting by leveraging cycle consistency across tim
e. We quantitatively show that our method improves label-propagation by a notewo
rthy margin of 13.1 mIoU on the ApolloScape dataset. Furthermore, by training wi
th the auto-labelled frames, we achieve competitive results on three semantic-se
gmentation benchmarks, improving the state-of-the-art by a large margin of 1.8 a
nd 3.61 mIoU on NYU-V2 and KITTI, while matching the current best results on Cit
yscapes.
****

ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models
Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, Sungroh Yoon; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 14367-14376
Denoising diffusion probabilistic models (DDPM) have shown remarkable performanc
e in unconditional image generation. However, due to the stochasticity of the ge
nerative process in DDPM, it is challenging to generate images with the desired
semantics. In this work, we propose Iterative Latent Variable Refinement (ILVR),
 a method to guide the generative process in DDPM to generate high-quality image
s based on a given reference image. Here, the refinement of the generative proce
ss in DDPM enables a single DDPM to sample images from various sets directed by
the reference image. The proposed ILVR method generates high-quality images whil
e controlling the generation. The controllability of our method allows adaptatio
n of a single DDPM without any additional learning in various image generation t
asks, such as generation from various downsampling factors, multi-domain image t

ranslation, paint-to-image, and editing with scribbles.
****

STVGBert: A Visual-Linguistic Transformer Based Framework for Spatio-Temporal Video Grounding
Rui Su, Qian Yu, Dong Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1533-1542
Spatio-temporal video grounding (STVG) aims to localize a spatio-temporal tube of a target object in an untrimmed video based on a query sentence. In this work, we propose a one-stage visual-linguistic transformer based framework called STVGBert for the STVG task, which can simultaneously localize the target object in both spatial and temporal domains. Specifically, without resorting to pre-generated object proposals, our STVGBert directly takes a video and a query sentence as the input, and then produces the cross-modal features by using the newly introduced cross-modal feature learning module ST-ViLBert. Based on the cross-modal features, our method then generates bounding boxes and predicts the starting and ending frames to produce the predicted object tube. To the best of our knowledge, our STVGBert is the first one-stage method, which can handle the STVG task without relying on any pre-trained object detectors. Comprehensive experiments demonstrate our newly proposed framework outperforms the state-of-the-art multi-stage methods on two benchmark datasets Vid-STG and HC-STVG.
****

Universal Representation Learning From Multiple Domains for Few-Shot Classification
Wei-Hong Li, Xialei Liu, Hakan Bilen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9526-9535
In this paper, we look at the problem of few-shot image classification that aims to learn a classifier for previously unseen classes and domains from few labeled samples. Recent methods use various adaptation strategies for aligning their visual representations to new domains or select the relevant ones from multiple domain-specific feature extractors. In this work, we present URL, which learns a single set of universal visual representations by distilling knowledge of multiple domain-specific networks after co-aligning their features with the help of adapters and centered kernel alignment. We show that the universal representations can be further refined for previously unseen domains by an efficient adaptation step in a similar spirit to distance learning methods. We rigorously evaluate our model in the recent Meta-Dataset benchmark and demonstrate that it significantly outperforms the previous methods while being more efficient.
****

Pseudo-Loss Confidence Metric for Semi-Supervised Few-Shot Learning
Kai Huang, Jie Geng, Wen Jiang, Xinyang Deng, Zhe Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8671-8680
Semi-supervised few-shot learning is developed to train a classifier that can adapt to new tasks with limited labeled data and a fixed quantity of unlabeled data. Most semi-supervised few-shot learning methods select pseudo-labeled data of unlabeled set by task-specific confidence estimation. This work presents a task-unified confidence estimation approach for semi-supervised few-shot learning, named pseudo-loss confidence metric (PLCM). It measures the data credibility by the loss distribution of pseudo-labels, which is synthetical considered multi-tasks. Specifically, pseudo-labeled data of different tasks are mapped to a unified metric space by mean of the pseudo-loss model, making it possible to learn the prior pseudo-loss distribution. Then, confidence of pseudo-labeled data is estimated according to the distribution component confidence of its pseudo-loss. Thus highly reliable pseudo-labeled data are selected to strengthen the classifier. Moreover, to overcome the pseudo-loss distribution shift and improve the effectiveness of classifier, we advance the multi-step training strategy coordinated with the class balance measures of class-apart selection and class weight. Experimental results on four popular benchmark datasets demonstrate that the proposed approach can effectively select pseudo-labeled data and achieve the state-of-the-art performance.
****

Learning Dual Priors for JPEG Compression Artifacts Removal
Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4086-4095
Deep learning (DL)-based methods have achieved great success in solving the ill-posed JPEG compression artifacts removal problem. However, as most DL architectures are designed to directly learn pixel-level mapping relationships, they largely ignore semantic-level information and lack sufficient interpretability. To address the above issues, in this work, we propose an interpretable deep network to learn both pixel-level regressive prior and semantic-level discriminative prior. Specifically, we design a variational model to formulate the image de-blocking problem and propose two prior terms for the image content and gradient, respectively. The content-relevant prior is formulated as a DL-based image-to-image regressor to perform as a de-blocker from the pixel-level. The gradient-relevant prior serves as a DL-based classifier to distinguish whether the image is compressed from the semantic-level. To effectively solve the variational model, we design an alternating minimization algorithm and unfold it into a deep network architecture. In this way, not only the interpretability of the deep network is increased, but also the dual priors can be well estimated from training samples. By integrating the two priors into a single framework, the image de-blocking problem can be well-constrained, leading to a better performance. Experiments on benchmarks and real-world use cases demonstrate the superiority of our method to the existing state-of-the-art approaches.
****

Searching for Two-Stream Models in Multivariate Space for Video Recognition
Xinyu Gong, Heng Wang, Mike Zheng Shou, Matt Feiszli, Zhangyang Wang, Zhicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8033-8042
Conventional video models rely on a single stream to capture the complex spatial-temporal features. Recent work on two-stream video models, such as SlowFast network and AssembleNet, prescribe separate streams to learn complementary features, and achieve stronger performance. However, manually designing both streams as well as the in-between fusion blocks is a daunting task, requiring to explore a tremendously large design space. Such manual exploration is time-consuming and often ends up with sub-optimal architectures when computational resources are limited and the exploration is insufficient. In this work, we present a pragmatic neural architecture search approach, which is able to search for two-stream video models in giant spaces efficiently. We design a multivariate search space, including 6 search variables to capture a wide variety of choices in designing two-stream models. Furthermore, we propose a progressive search procedure, by searching for the architecture of individual streams, fusion blocks and attention blocks one after the other. We demonstrate two-stream models with significantly better performance can be automatically discovered in our design space. Our searched two-stream models, namely Auto-TSNet, consistently outperform other models on standard benchmarks. On Kinetics, compared with the SlowFast model, our Auto-TSNet-L model reduces FLOPS by nearly 11 times while achieving the same accuracy 78.9%. On Something-Something-V2, Auto-TSNet-M improves the accuracy by at least 2% over other methods which use less than 50 GFLOPS per video.
****

Refining Activation Downsampling With SoftPool
Alexandros Stergiou, Ronald Poppe, Grigorios Kalliatakis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10357-10366
Convolutional Neural Networks (CNNs) use pooling to decrease the size of activation maps. This process is crucial to increase the receptive fields and to reduce computational requirements of subsequent convolutions. An important feature of the pooling operation is the minimization of information loss, with respect to the initial activation maps, without a significant impact on the computation and memory overhead. To meet these requirements, we propose SoftPool: a fast and efficient method for exponentially weighted activation downsampling. Through experiments across a range of architectures and pooling methods, we demonstrate that SoftPool can retain more information in the reduced activation maps. This refined

downsampling leads to improvements in a CNN's classification accuracy. Experiments with pooling layer substitutions on ImageNet1K show an increase in accuracy over both original architectures and other pooling methods. We also test SoftPool on video datasets for action recognition. Again, through the direct replacement of pooling layers, we observe consistent performance improvements while computational loads and memory requirements remain limited.

****

Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing

Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, Gerard Pons-Moll; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11708-11718

We present Neural Generalized Implicit Functions(Neural-GIF), to animate people in clothing as a function of the body pose. Given a sequence of scans of a subject in various poses, we learn to animate the character for new poses. Existing methods have relied on template-based representations of the human body(or clothing). However such models usually have fixed and limited resolutions, and require difficult data pre-processing steps, and cannot be used for complex clothing. We draw inspiration from template-based methods, which factorize motion into articulation and non-rigid deformation, but generalize this concept for implicit shape learning to obtain a more flexible model. We learn to map every point in the space to a canonical space, where a learned deformation field is applied to model non-rigid effects, before evaluating the signed distance field. Our formulation allows the learning of complex and non-rigid deformations of clothing and soft tissue, without computing a template registration as it is common with current approaches. Neural-GIF can be trained on raw 3D scans and reconstructs detailed complex surface geometry and deformations. Moreover, the model can generalize to new poses. We evaluate our method on a variety of characters from different public datasets in diverse clothing styles and show significant improvement over baseline methods, quantitatively and qualitatively. We also extend our model to multiple shape setting. To stimulate further research, we will make the model, code, and data publicly available.

****

Learning Multiple Pixelwise Tasks Based on Loss Scale Balancing

Jae-Han Lee, Chul Lee, Chang-Su Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5107-5116

We propose a novel loss weighting algorithm, called loss scale balancing (LSB), for multi-task learning (MTL) of pixelwise vision tasks. An MTL model is trained to estimate multiple pixelwise predictions using an overall loss, which is a linear combination of individual task losses. The proposed algorithm dynamically adjusts the linear weights to learn all tasks effectively. Instead of controlling the trend of each loss value directly, we balance the loss scale --- the product of the loss value and its weight --- periodically. In addition, by evaluating the difficulty of each task based on the previous loss record, the proposed algorithm focuses more on difficult tasks during training. Experimental results show that the proposed algorithm outperforms conventional weighting algorithms for MTL of various pixelwise tasks. Codes are available at https://github.com/jaehanlee-mcl/LSB-MTL.

****

MLVSNet: Multi-Level Voting Siamese Network for 3D Visual Tracking

Zhoutao Wang, Qian Xie, Yu-Kun Lai, Jing Wu, Kun Long, Jun Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3101-3110

Benefiting from the excellent performance of Siamese-based trackers, huge progress on 2D visual tracking has been achieved. However, 3D visual tracking is still under-explored. Inspired by the idea of Hough voting in 3D object detection, in this paper, we propose a Multi-level Voting Siamese Network (MLVSNet) for 3D visual tracking from outdoor point cloud sequences. To deal with sparsity in outdoor 3D point clouds, we propose to perform Hough voting on multi-level features to get more vote centers and retain more useful information, instead of voting on

ly on the final level feature as in previous methods. We also design an efficien t and lightweight Target-Guided Attention (TGA) module to transfer the target in formation and highlight the target points in the search area. Moreover, we propo se a Vote-cluster Feature Enhancement (VFE) module to exploit the relationships between different vote clusters. Extensive experiments on the 3D tracking benchm ark of KITTI dataset demonstrate that our MLVSNet outperforms state-of-the-art m ethods with significant margins. Code will be available at https://github.com/Co deWZT/MLVSNet.
****

ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot
Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2021, pp. 112-121
One-stage long-tailed recognition methods improve the overall performance in a " seesaw" manner, i.e., either sacrifice the head's accuracy for better tail class ification or elevate the head's accuracy even higher but ignore the tail. Existi ng algorithms bypass such trade-off by a multi-stage training process: pre-train ing on imbalanced set and fine-tuning on balanced set. Though achieving promisin g performance, not only are they sensitive to the generalizability of the pre-tr ained model, but also not easily integrated into other computer vision tasks lik e detection and segmentation, where pre-training of classifier solely is not app licable. In this paper, we propose a one-stage long-tailed recognition scheme, a lly complementary experts (ACE), where the expert is the most knowledgeable spec ialist in a sub-set that dominates its training, and is complementary to other e xperts in the less-seen categories without disturbed by what it has never seen. We design a distribution-adaptive optimizer to adjust the learning pace of each expert to avoid over-fitting. Without special bells and whistles, the vanilla AC E outperforms the current one-stage SOTA method by 3 10% on CIFAR10-LT, CIFAR100 -LT, ImageNet-LT and iNaturalist datasets. It is also shown to be the first one to break the "seesaw" trade-off by improving the accuracy of the majority and mi nority categories simultaneously in only one stage.
****

Hyperspectral Image Denoising With Realistic Data
Tao Zhang, Ying Fu, Cheng Li; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2021, pp. 2248-2257
The hyperspectral image (HSI) denoising has been widely utilized to improve HSI qualities. Recently, learning-based HSI denoising methods have shown their effec tiveness, but most of them are based on synthetic dataset and lack the generaliz ation capability on real testing HSI. Moreover, there is still no public paired real HSI denoising dataset to learn HSI denoising network and quantitatively eva luate HSI methods. In this paper, we mainly focus on how to produce realistic da taset for learning and evaluating HSI denoising network. On the one hand, we col lect a paired real HSI denoising dataset, which consists of shortexposure noisy HSIs and the corresponding long-exposure clean HSIs. On the other hand, we propo se an accurate HSI noise model which matches the distribution of real data well and can be employed to synthesize realistic dataset. On the basis of the noise m odel, we present an approach to calibrate the noise parameters of the given hype rspectral camera. The extensive experimental results show that a network learned with only synthetic data generated by our noise model performs as well as it is learned with paired real data.
****

Collaborative Learning With Disentangled Features for Zero-Shot Domain Adaptatio n
Won Young Jhoo, Jae-Pil Heo; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2021, pp. 8896-8905
Typical domain adaptation techniques aim to transfer label information from a la bel-rich source domain to a label-scarce target domain in the same label space. However, it is often hard to get even the unlabeled target domain data of a task of interest. In such a case, we can capture the domain shift between the source domain and target domain from an unseen task and transfer it to the task of int erest, which is known as zero-shot domain adaptation (ZSDA). Existing state-of-t

he-art methods for ZSDA attempted to generate target domain data. However, train ing such generative models causes significant computational overhead and is hard ly optimized. In this paper, we propose a novel ZSDA method that learns a task-a gnostic domain shift by collaborative training of domain-invariant semantic feat ures and task-invariant domain features via adversarial learning. Meanwhile, the spatial attention map is learned from disentangled feature representations to s electively emphasize the domain-specific salient parts of the domain-invariant f eatures. Experimental results show that our ZSDA method achieves state-of-the-ar t performance on several benchmarks.
****

Rethinking Noise Synthesis and Modeling in Raw Denoising
Yi Zhang, Hongwei Qin, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4593-4601
The lack of large-scale real raw image denoising dataset gives the rise to chall enges on synthesizing realistic raw image noise for training denoising models. H owever, the real raw image noise is contributed by many noise sources and varies greatly among different sensors. Existing methods are unable to model all noise sources accurately, and building a noise model for each sensor is also laboriou s. In this paper, we introduce a new perspective to synthesize noise by directly sampling from the sensor's real noise. It inherently generates accurate raw ima ge noise for different camera sensors. Two efficient and generic techniques: pat tern-aligned patch sampling and high-bit reconstruction help accurate synthesis of spatial-correlated noise and high-bit noise respectively. We conduct systemat ic experiments on SIDD and ELD datasets. The results show that (1) our method ou tperforms existing methods and demonstrates wide generalization on different sen sors and lighting conditions. (2) Recent conclusions derived from DNN-based nois e modeling methods are actually based on inaccurate noise parameters. The DNN-ba sed methods still cannot outperform physics-based statistical methods.
****

Disentangled Representation for Age-Invariant Face Recognition: A Mutual Informa tion Minimization Perspective
Xuege Hou, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 3692-3701
General face recognition has seen remarkable progress in recent years. However, large age gap still remains a big challenge due to significant alterations in fa cial appearance and bone structure. Disentanglement plays a key role in partitio ning face representations into identity-dependent and age-dependent components f or age-invariant face recognition (AIFR). In this paper we propose a multi-task learning framework based on mutual information minimization (MT-MIM), which cast s the disentangled representation learning as an objective of information constr aints. The method trains a disentanglement network to minimize mutual informatio n between the identity component and age component of the face image from the sa me person, and reduce the effect of age variations during the identification pro cess. For quantitative measure of the degree of disentanglement, we verify that mutual information can represent as metric. The resulting identity-dependent rep resentations are used for age-invariant face recognition. We evaluate MT-MIM on popular public-domain face aging datasets (FG-NET, MORPH Album 2, CACD and AgeDB) and obtained significant improvements over previous state-of-the-art methods. Specifically, our method exceeds the baseline models by over 0.4% on MORPH Album 2, and over 0.7% on CACD subsets, which are impressive improvements at the high accuracy levels of above 99% and an average of 94%.
****

Contact-Aware Retargeting of Skinned Motion
Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, Jun Saito; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9720-9729
This paper introduces a motion retargeting method that preserves self-contacts a nd prevents inter-penetration. Self-contacts, such as when hands touch each othe r or the torso or the head, are important attributes of human body language and dynamics, yet existing methods do not model or preserve these contacts. Likewise

, self-penetrations, such as a hand passing into the torso, are a typical artifa
ct of motion estimation methods. The input to our method is a human motion seque
nce and a target skeleton and character geometry. The method identifies self-con
tacts and ground contacts in the input motion, and optimizes the motion to apply
 to the output skeleton, while preserving these contacts and reducing self-penet
rations. We introduce a novel geometry-conditioned recurrent network with an enc
oder-space optimization strategy that achieves efficient retargeting while satis
fying contact constraints. In experiments, our results quantitatively outperform
 previous methods and in the user study our retargeted motions are rated as high
er-quality than those produced by recent works. We also show our method generali
zes to motion estimated from human videos where we improve over previous works t
hat produce noticeable interpenetration.
****

Box-Aware Feature Enhancement for Single Object Tracking on Point Clouds
Chaoda Zheng, Xu Yan, Jiantao Gao, Weibing Zhao, Wei Zhang, Zhen Li, Shuguang Cu
i; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2021, pp. 13199-13208
Current 3D single object tracking approaches track the target based on a feature
 comparison between the target template and the search area. However, due to the
 common occlusion in LiDAR scans, it is non-trivial to conduct accurate feature
comparisons on severe sparse and incomplete shapes. In this work, we exploit the
 ground truth bounding box given in the first frame as a strong cue to enhance t
he feature description of the target object, enabling a more accurate feature co
mparison in a simple yet effective way. In particular, we first propose the BoxC
loud, an informative and robust representation, to depict an object using the po
int-to-box relation. We further design an efficient box-aware feature fusion mod
ule, which leverages the aforementioned BoxCloud for reliable feature matching a
nd embedding. Integrating the proposed general components into an existing model
 P2B, we construct a superior box-aware tracker (BAT). Experiments confirm that
our proposed BAT outperforms the previous state-of-the-art by a large margin on
both KITTI and NuScenes benchmarks, achieving a 12.8% improvement in terms of pr
ecision while running 20% faster.
****

FATNN: Fast and Accurate Ternary Neural Networks
Peng Chen, Bohan Zhuang, Chunhua Shen; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2021, pp. 5219-5228
Ternary Neural Networks (TNNs) have received much attention due to being potenti
ally orders of magnitude faster in inference, as well as more power efficient, t
han full-precision counterparts. However, 2 bits are required to encode the tern
ary representation with only 3 quantization levels leveraged. As a result, conve
ntional TNNs have similar memory consumption and speed compared with the standar
d 2-bit models, but have worse representational capability. Moreover, there is s
till a significant gap in accuracy between TNNs and full-precision networks, ham
pering their deployment to real applications. To tackle these two challenges, in
 this work, we first show that, under some mild constraints, computational compl
exity of the ternary inner product can be reduced by 2x. Second, to mitigate the
 performance gap, we elaborately design an implementation-dependent ternary quan
tization algorithm. The proposed framework is termed Fast and Accurate Ternary N
eural Networks (FATNN). Experiments on image classification demonstrate that our
 FATNN surpasses the state-of-the-arts by a significant margin in accuracy. More
 importantly, speedup evaluation compared with various precisions is analyzed on
 several platforms, which serves as a strong benchmark for further research.
****

Multitask AET With Orthogonal Tangent Regularity for Dark Object Detection
Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, Tatsuya Harada; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
 pp. 2553-2562
Dark environment becomes a challenge for computer vision algorithms owing to ins
ufficient photons and undesirable noises. Most of the existing studies tackle th
is by either targeting human vision for better visual perception or improving th

e machine vision for specific high-level tasks. In addition, these methods rely on data argumentation and directly train their models based on real-world or over-simplified synthetic datasets without exploring the intrinsic pattern behind illumination translation. Here, we propose a novel multitask auto encoding transformation (MAET) model that combines human vision and machine vision tasks to enhance object detection in a dark environment. With a self-supervision learning, the MAET learns an intrinsic visual structure by encoding and decoding the realistic illumination-degrading transformation considering the physical noise model and image signal processing (ISP). Based on this representation, we achieve object detection task by decoding the bounding box coordinates and classes. To avoid the over-entanglement of two tasks, our MAET disentangles the object and degrading features by imposing an orthogonal tangent regularity. This forms a parametric manifold along which multitask predictions can be geometrically formulated by maximizing the orthogonality between the tangents along the outputs of respective tasks. Our framework can be implemented based on the mainstream object detection architecture and directly trained end-to-end using the normal target detection datasets, such as COCO and VOC. We have achieved the state-of-the-art performance using synthetic and real-world datasets.

****

Field-Guide-Inspired Zero-Shot Learning

Utkarsh Mall, Bharath Hariharan, Kavita Bala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9546-9555

Modern recognition systems require large amounts of supervision to achieve accuracy. Adapting to new domains requires significant data from experts, which is onerous and can become too expensive. Zero-shot learning requires an annotated set of attributes for a novel category. Annotating the full set of attributes for a novel category proves to be a tedious and expensive task in deployment. This is especially the case when the recognition domain is an expert domain. We introduce a new field-guide-inspired approach to zero-shot annotation where the learner model interactively asks for the most useful attributes that define a class. We evaluate our method on classification benchmarks with attribute annotations like CUB, SUN, and AWA2 and show that our model achieves the performance of a model with full annotations at the cost of a significantly fewer number of annotations. Since the time of experts is precious, decreasing annotation cost can be very valuable for real-world deployment.

****

Contrastive Attention Maps for Self-Supervised Co-Localization

Minsong Ki, Youngjung Uh, Junsuk Choe, Hyeran Byun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2803-2812

The goal of unsupervised co-localization is to locate the object in a scene under the assumptions that 1) the dataset consists of only one superclass, e.g., birds, and 2) there are no human-annotated labels in the dataset. The most recent method achieves impressive co-localization performance by employing self-supervised representation learning approaches such as predicting rotation. In this paper, we introduce a new contrastive objective directly on the attention maps to enhance co-localization performance. Our contrastive loss function exploits rich information of location, which induces the model to activate the extent of the object effectively. In addition, we propose a pixel-wise attention pooling that selectively aggregates the feature map regarding their magnitudes across channels. Our methods are simple and shown effective by extensive qualitative and quantitative evaluation, achieving state-of-the-art co-localization performances by large margins on four datasets: CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Stanford Dogs. Our code will be publicly available online for the research community.

****

ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, Simone Stumpf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10818-10828

Object recognition has made great advances in the last decade, but predominately still relies on many high-quality training examples per object category. In contrast, learning new objects from only a few examples could enable many impactful applications from robotics to user personalization. Most few-shot learning research, however, has been driven by benchmark datasets that lack the high variation that these applications will face when deployed in the real-world. To close this gap, we present the ORBIT dataset and benchmark, grounded in the real-world application of teachable object recognizers for people who are blind/low-vision. The dataset contains 3,822 videos of 486 objects recorded by people who are blind/low-vision on their mobile phones. The benchmark reflects a realistic, highly challenging recognition problem, providing a rich playground to drive research in robustness to few-shot, high-variation conditions. We set the benchmark's first state-of-the-art and show there is massive scope for further innovation, holding the potential to impact a broad range of real-world vision applications including tools for the blind/low-vision community. We release the dataset at https://doi.org/10.25383/city.14294597 and benchmark code at https://github.com/microsoft/ORBIT-Dataset.

****

Domain Generalization via Gradient Surgery

Lucas Mansilla, Rodrigo Echeveste, Diego H. Milone, Enzo Ferrante; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6630-6638

In real-life applications, machine learning models often face scenarios where there is a change in data distribution between training and test domains. When the aim is to make predictions on distributions different from those seen at training, we incur in a domain generalization problem. Methods to address this issue learn a model using data from multiple source domains, and then apply this model to the unseen target domain. Our hypothesis is that when training with multiple domains, conflicting gradients within each mini-batch contain information specific to the individual domains which is irrelevant to the others, including the test domain. If left untouched, such disagreement may degrade generalization performance. In this work, we characterize the conflicting gradients emerging in domain shift scenarios and devise novel gradient agreement strategies based on gradient surgery to alleviate their effect. We validate our approach in image classification tasks with three multi-domain datasets, showing the value of the proposed agreement strategy in enhancing the generalization capability of deep learning models in domain shift scenarios.

****

Semi-Supervised Single-Stage Controllable GANs for Conditional Fine-Grained Image Generation

Tianyi Chen, Yi Liu, Yunfei Zhang, Si Wu, Yong Xu, Feng Liangbing, Hau San Wong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9264-9273

Previous state-of-the-art deep generative models improve fine-grained image generation quality by designing hierarchical model structures and synthesizing images across multiple stages. The learning process is typically performed without any supervision in object categories. To address this issue, while at the same time to alleviate the level of complexity of both model design and training, we propose a Single-Stage Controllable GAN (SSC-GAN) for conditional fine-grained image synthesis in a semi-supervised setting. Considering the fact that fine-grained object categories may have subtle distinctions and shared attributes, we take into account three factors of variation for generative modeling: class-independent content, cross-class attributes and class semantics, and associate them with different variables. To ensure disentanglement among the variables, we maximize mutual information between the class-independent variable and synthesized images, map real images to the latent space of a generator to perform consistency regularization of cross-class attributes, and incorporate class semantic-based regularization into a discriminator's feature space. We show that the proposed approach delivers a single-stage controllable generator and high-fidelity synthesized images of fine-grained categories. The proposed approach establishes state-of-the

-art semi-supervised image synthesis results across multiple fine-grained datase
ts.
****

Partner-Assisted Learning for Few-Shot Image Classification
Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, Wael Abd-Al
mageed; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 10573-10582
Few-shot Learning has been studied to mimic human visual capabilities and learn
effective models without the need of exhaustive human annotation. Even though th
e idea of meta-learning for adaptation has dominated the few-shot learning metho
ds, how to train a feature extractor is still a challenge. In this paper, we foc
us on the design of training strategy to obtain an elemental representation such
 that the prototype of each novel class can be estimated from a few labeled samp
les. We propose a two-stage training scheme, Partner-Assisted Learning (PAL), wh
ich first trains a partner encoder to model pair-wise similarities and extract f
eatures serving as soft-anchors, and then trains a main encoder by aligning its
outputs with soft-anchors while attempting to maximize classification performanc
e. Two alignment constraints from logit-level and feature-level are designed ind
ividually. For each few-shot task, we perform prototype classification. Our meth
od consistently outperforms the state-of-the-art method on four benchmarks. Deta
iled ablation studies of PAL are provided to justify the selection of each compo
nent involved in training.
****

Contrastive Coding for Active Learning Under Class Distribution Mismatch
Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, Cuiping Li; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8927
-8936
Active learning (AL) is successful based on the assumption that labeled and unla
beled data are obtained from the same class distribution. However, its performan
ce deteriorates under class distribution mismatch, wherein the unlabeled data co
ntain many samples out of the class distribution of labeled data. To effectively
 handle the problems under class distribution mismatch, we propose a contrastive
 coding based AL framework named CCAL. Unlike the existing AL methods that focus
 on selecting the most informative samples for annotating, CCAL extracts both se
mantic and distinctive features by contrastive learning and combines them in a q
uery strategy to choose the most informative unlabeled samples with matched cate
gories. Theoretically, we prove that the AL error of CCAL has a tight upper boun
d. Experimentally, we evaluate its performance on CIFAR10, CIFAR100, and an arti
ficial cross-dataset that consists of five datasets; consequently, CCAL achieves
 state-of-the-art performance by a large margin with remarkably lower annotation
 cost. To the best of our knowledge, CCAL is the first work related to AL for cl
ass distribution mismatch.
****

Partial Video Domain Adaptation With Partial Adversarial Temporal Attentive Netw
ork
Yuecong Xu, Jianfei Yang, Haozhi Cao, Zhenghua Chen, Qi Li, Kezhi Mao; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
 9332-9341
Partial Domain Adaptation (PDA) is a practical and general domain adaptation sce
nario, which relaxes the fully shared label space assumption such that the sourc
e label space subsumes the target one. The key challenge of PDA is the issue of
negative transfer caused by source-only classes. For videos, such negative trans
fer could be triggered by both spatial and temporal features, which leads to a m
ore challenging Partial Video Domain Adaptation (PVDA) problem. In this paper, w
e propose a novel Partial Adversarial Temporal Attentive Network (PATAN) to addr
ess the PVDA problem by utilizing both spatial and temporal features for filteri
ng source-only classes. Besides, PATAN constructs effective overall temporal fea
tures by attending to local temporal features that contribute more toward the cl
ass filtration process. We further introduce new benchmarks to facilitate resear
ch on PVDA problems, covering a wide range of PVDA scenarios. Empirical results

demonstrate the state-of-the-art performance of our proposed PATAN across the multiple PVDA benchmarks.
****

Personalized and Invertible Face De-Identification by Disentangled Identity Information Manipulation
Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, Li Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3334-3342
The popularization of intelligent devices including smartphones and surveillance cameras results in more serious privacy issues. De-identification is regarded as an effective tool for visual privacy protection with the process of concealing or replacing identity information. Most of the existing de-identification methods suffer from some limitations since they mainly focus on the protection process and are usually non-reversible. In this paper, we propose a personalized and invertible de-identification method based on the deep generative model, where the main idea is introducing a user-specific password and an adjustable parameter to control the direction and degree of identity variation. Extensive experiments demonstrate the effectiveness and generalization of our proposed framework for both face de-identification and recovery.
****

Learning Compatible Embeddings
Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, Feng Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9939-9948
Achieving backward compatibility when rolling out new models can highly reduce costs or even bypass feature re-encoding of existing gallery images for in-production visual retrieval systems. Previous related works usually leverage losses used in knowledge distillation which can cause performance degradations or not guarantee compatibility. To address these issues, we propose a general framework called Learning Compatible Embeddings (LCE) which is applicable for both cross model compatibility and compatible training in direct/forward/backward manners. Our compatibility is achieved by aligning class centers between models directly or via a transformation, and restricting more compact intra-class distributions for the new model. Experiments are conducted in extensive scenarios such as changes of training dataset, loss functions, network architectures as well as feature dimensions, and demonstrate that LCE efficiently enables model compatibility with marginal sacrifices of accuracies.
****

Seasonal Contrast: Unsupervised Pre-Training From Uncurated Remote Sensing Data
Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vazquez, Pau Rodríguez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9414-9423
Remote sensing and automatic earth monitoring are key to solve global-scale challenges such as disaster prevention, land use monitoring, or tackling climate change. Although there exist vast amounts of remote sensing data, most of it remains unlabeled and thus inaccessible for supervised learning algorithms. Transfer learning approaches can reduce the data requirements of deep learning algorithms. However, most of these methods are pre-trained on ImageNet and their generalization to remote sensing imagery is not guaranteed due to the domain gap. In this work, we propose Seasonal Contrast (SeCo), an effective pipeline to leverage unlabeled data for in-domain pre-training of remote sensing representations. The SeCo pipeline is composed of two parts. First, a principled procedure to gather large-scale, unlabeled and uncurated remote sensing datasets containing images from multiple Earth locations at different timestamps. Second, a self-supervised algorithm that takes advantage of time and position invariance to learn transferable representations for remote sensing applications. We empirically show that models trained with SeCo achieve better performance than their ImageNet pre-trained counterparts and state-of-the-art self-supervised learning methods on multiple downstream tasks. The datasets and models in SeCo will be made public to facilitate transfer learning and enable rapid progress in remote sensing applications.
****

Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation

Zechen Bai, Yuta Nakashima, Noa Garcia; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2021, pp. 5422-5432
Have you ever looked at a painting and wondered what is the story behind it? Thi
s work presents a framework to bring art closer to people by generating comprehe
nsive descriptions of fine-art paintings. Generating informative descriptions fo
r artworks, however, is extremely challenging, as it requires to 1) describe mul
tiple aspects of the image such as its style, content, or composition, and 2) pr
ovide background and contextual knowledge about the artist, their influences, or
 the historical period. To address these challenges, we introduce a multi-topic
and knowledgeable art description framework, which modules the generated sentenc
es according to three artistic topics and, additionally, enhances each descripti
on with external knowledge. The framework is validated through an exhaustive ana
lysis, both quantitative and qualitative, as well as a comparative human evaluat
ion, demonstrating outstanding results in terms of both topic diversity and info
rmation veracity.
****

Unsupervised Curriculum Domain Adaptation for No-Reference Video Quality Assessm
ent
Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, Guangming Shi; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5178
-5187
During the last years, convolutional neural networks (CNNs) have triumphed over
video quality assessment (VQA) tasks. However, CNN-based approaches heavily rely
 on annotated data which are typically not available in VQA, leading to the diff
iculty of model generalization. Recent advances in domain adaptation technique m
akes it possible to adapt models trained on source data to unlabeled target data
. However, due to the distortion diversity and content variation of the collecte
d videos, the intrinsic subjectivity of VQA tasks hampers the adaptation perform
ance. In this work, we propose a curriculum-style unsupervised domain adaptation
 to handle the cross-domain no-reference VQA problem. The proposed approach coul
d be divided into two stages. In the first stage, we conduct an adaptation betwe
en source and target domains to predict the rating distribution for target sampl
es, which can better reveal the subjective nature of VQA. From this adaptation,
we split the data in target domain into confident and uncertain subdomains using
 the proposed uncertainty-based ranking function, through measuring their predic
tion confidences. In the second stage, by regarding samples in confident subdoma
in as the easy tasks in the curriculum, a fine-level adaptation is conducted bet
ween two subdomains to fine-tune the prediction model. Extensive experimental re
sults on benchmark datasets highlight the superiority of the proposed method ove
r the competing methods in both accuracy and speed. The source code is released
at https://github.com/cpf0079/UCDA.
****

GTT-Net: Learned Generalized Trajectory Triangulation
Xiangyu Xu, Enrique Dunn; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2021, pp. 5795-5804
We present GTT-Net, a supervised learning framework for the reconstruction of sp
arse dynamic 3D geometry. We build on a graph-theoretic formulation of the gener
alized trajectory triangulation problem, where non-concurrent multi-view imaging
 geometry is known but global image sequencing is not provided. GTT-Net learns p
airwise affinities modeling the spatio-temporal relationships among our input ob
servations and leverages them to determine 3D geometry estimates. Experiments re
constructing 3D motion-capture sequences show GTT-Net outperforms the state of t
he art in terms of accuracy and robustness. Within the context of articulated mo
tion reconstruction, our proposed architecture is 1) able to learn and enforce s
emantic 3D motion priors for shared training and test domains, while being 2) ab
le to generalize its performance across different training and test domains. Mor
eover, GTT-Net provides a computationally streamlined framework for trajectory t
riangulation with applications to multi-instance reconstruction and event segmen
tation.
****

Contrastive Learning of Image Representations With Cross-Video Cycle-Consistency
Haiping Wu, Xiaolong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10149-10159
Recent works have advanced the performance of self-supervised representation learning by a large margin. The core among these methods is intra-image invariance learning. Two different transformations of one image instance are considered as a positive sample pair, where various tasks are designed to learn invariant representations by comparing the pair. Analogically, for video data, representations of frames from the same video are trained to be closer than frames from other videos, i.e. intra-video invariance. However, cross-video relation has barely been explored for visual representation learning. Unlike intra-video invariance, ground-truth labels of cross-video relation is usually unavailable without human labors. In this paper, we propose a novel contrastive learning method which explores the cross-video relation by using cycle-consistency for general image representation learning. This allows to collect positive sample pairs across different video instances, which we hypothesize will lead to higher-level semantics. We validate our method by transferring our image representation to multiple downstream tasks including visual object tracking, image classification, and action recognition. We show significant improvement over state-of-the-art contrastive learning methods. Project page is available at https://happywu.github.io/cycle_contrast_video.
****
Learning To Remove Refractive Distortions From Underwater Images
Simron Thapa, Nianyi Li, Jinwei Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5007-5016
The fluctuation of the water surface causes refractive distortions that severely downgrade the image of an underwater scene. Here, we present the distortion-guided network (DG-Net) for restoring distortion-free underwater images. The key idea is to use a distortion map to guide network training. The distortion map models the pixel displacement caused by water refraction. We first use a physically constrained convolutional network to estimate the distortion map from the refracted image. We then use a generative adversarial network guided by the distortion map to restore the sharp distortion-free image. Since the distortion map indicates correspondences between the distorted image and the distortion-free one, it guides the network to make better predictions. We evaluate our network on several real and synthetic underwater image datasets and show that it out-performs the state-of-the-art algorithms, especially in presence of large distortions. We also show results of complex scenarios, including outdoor swimming pool images captured by the drone and indoor aquarium images taken by cellphone camera.
****
BARF: Bundle-Adjusting Neural Radiance Fields
Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, Simon Lucey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5741-5751
Neural Radiance Fields (NeRF) have recently gained a surge of interest within the computer vision community for its power to synthesize photorealistic novel views of real-world scenes. One limitation of NeRF, however, is its requirement of known camera poses to learn the scene representations. In this paper, we propose Bundle-Adjusting Neural Radiance Fields (BARF) for training NeRF from imperfect camera poses -- the joint problem of learning neural 3D representations and registering camera frames. We establish a theoretical connection to classical planar image registration and show that coarse-to-fine registration is also applicable to NeRF. Furthermore, we demonstrate mathematically that positional encoding has a direct impact on the basin of attraction for registration with a synthesis-based objective. Experiments on synthetic and real-world data show that BARF can effectively optimize the neural scene representations and resolve large camera pose misalignment at the same time. This enables applications of view synthesis and localization of video sequences from unknown camera poses, opening up new avenues for visual localization systems (e.g. SLAM) towards sequential registration with NeRF.
****

Seeking Similarities Over Differences: Similarity-Based Domain Alignment for Adaptive Object Detection
Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, Bernt Schiele; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9204-9213
In order to robustly deploy object detectors across a wide range of scenarios, they should be adaptable to shifts in the input distribution without the need to constantly annotate new data. This has motivated research in Unsupervised Domain Adaptation (UDA) algorithms for detection. UDA methods learn to adapt from labeled source domains to unlabeled target domains, by inducing alignment between detector features from source and target domains. Yet, there is no consensus on what features to align and how to do the alignment. In our work, we propose a framework that generalizes the different components commonly used by UDA methods laying the ground for an in-depth analysis of the UDA design space. Specifically, we propose a novel UDA algorithm, ViSGA, a direct implementation of our framework, that leverages the best design choices and introduces a simple but effective method to aggregate features at the instance-level based on the visual similarity before inducing group alignment via adversarial training. We show that both similarity-based grouping and adversarial training allows our model to focus on coarsely aligning feature groups, without being forced to match all instances across loosely aligned domains. Finally, we examine the applicability of ViSGA to the setting where labeled data are gathered from different sources. Experiments show that not only our method outperforms previous single-source approaches on Sim2Real and Adverse Weather, but also generalizes well to the multi-source setting.
****

LapsCore: Language-Guided Person Search via Color Reasoning
Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, Shuguang Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1624-1633
The key point of language-guided person search is to construct the cross-modal association between visual and textual input. Existing methods focus on designing multimodal attention mechanisms and novel cross-modal loss functions to learn such association implicitly. We propose a representation learning method for language-guided person search based on color reasoning (LapsCore). It can explicitly build a fine-grained cross-modal association bidirectionally. Specifically, a pair of dual sub-tasks, image colorization and text completion, is designed. In the former task, rich text information is learned to colorize gray images, and the latter one requests the model to understand the image and complete color word vacancies in the captions. The two sub-tasks enable models to learn correct alignments between text phrases and image regions, so that rich multimodal representations can be learned. Extensive experiments on multiple datasets demonstrate the effectiveness and superiority of the proposed method.
****

Deep Permutation Equivariant Structure From Motion
Dror Moran, Hodaya Koslowsky, Yoni Kasten, Haggai Maron, Meirav Galun, Ronen Basri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5976-5986
Existing deep methods produce highly accurate 3D reconstructions in stereo and multiview stereo settings, i.e., when cameras are both internally and externally calibrated. Nevertheless, the challenge of simultaneous recovery of camera poses and 3D scene structure in multiview settings with deep networks is still outstanding. Inspired by projective factorization for Structure from Motion (SFM) and by deep matrix completion techniques, we propose a neural network architecture that, given a set of point tracks in multiple images of a static scene, recovers both the camera parameters and a (sparse) scene structure by minimizing an unsupervised reprojection loss. Our network architecture is designed to respect the structure of the problem: the sought output is equivariant to permutations of both cameras and scene points. Notably, our method does not require initialization of camera parameters or 3D point locations. We test our architecture in two setups: (1) single scene reconstruction and (2) learning from multiple scenes. Our e

xperiments, conducted on a variety of datasets in both internally calibrated and uncalibrated settings, indicate that our method accurately recovers pose and structure, on par with classical state of the art methods. Additionally, we show that a pre-trained network can be used to reconstruct novel scenes using inexpensive fine-tuning with no loss of accuracy.

****

Collaborative Unsupervised Visual Representation Learning From Decentralized Data

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, Shuai Yi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4912-4921

Unsupervised representation learning has achieved outstanding performances using centralized data available on the Internet. However, the increasing awareness of privacy protection limits sharing of decentralized unlabeled image data that grows explosively in multiple parties (e.g. mobile phones and cameras). As such, a natural problem is how to leverage these data to learn visual representations for downstream tasks while preserving data privacy. To address this problem, we propose a novel federated unsupervised learning framework, FedU. In this framework, each party trains models from unlabeled data independently using contrastive learning with an online network and a target network. Then, a central server aggregates trained models and updates clients' models with the aggregated global model. It preserves data privacy as each party only has access to its raw data. Decentralized data among multiple parties is normally non-independent and identically distributed (non-IID), which leads to performance degradation. To tackle this challenge, we propose two simple but effective methods: (1) We design the communication protocol to upload only the encoders of online networks for server aggregation and update them with the aggregated encoder. (2) We introduce a new module to dynamically decide how to update the predictors based on the degree of divergence caused by non-IID. The predictor is the other component of the online network. Extensive experiments and ablations demonstrate the effectiveness and significance of FedU. It outperforms training with only one party by over 5% and other methods by over 14% in linear and semi-supervised evaluation on non-IID data.

****

DeepPRO: Deep Partial Point Cloud Registration of Objects

Donghoon Lee, Onur C. Hamsici, Steven Feng, Prachee Sharma, Thorsten Gernoth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5683-5692

We consider the problem of online and real-time registration of partial point clouds obtained from an unseen real-world rigid object without knowing its 3D model. The point cloud is partial as it is obtained by a depth sensor capturing only the visible part of the object from a certain viewpoint. It introduces two main challenges: 1) two partial point clouds do not fully overlap and 2) keypoints tend to be less reliable when the visible part of the object does not have salient local structures. To address these issues, we propose DeepPRO, a keypoint-free and an end-to-end trainable deep neural network. Its core idea is inspired by how humans align two point clouds: we can imagine how two point clouds will look like after the registration based on their shape. To realize the idea, DeepPRO has inputs of two partial point clouds and directly predicts the point-wise location of the aligned point cloud. By preserving the ordering of points during the prediction, we enjoy dense correspondences between input and predicted point clouds when inferring rigid transform parameters. We conduct extensive experiments on the real-world Linemod and synthetic ModelNet40 datasets. In addition, we collect and evaluate on the PRO1k dataset, a large-scale version of Linemod meant to test generalization to real-world scans. Results show that DeepPRO achieves the best accuracy against thirteen strong baseline methods, e.g., 2.2mm ADD on the Linemod dataset, while running 50 fps on mobile devices.

****

RECALL: Replay-Based Continual Learning in Semantic Segmentation

Andrea Maracani, Umberto Michieli, Marco Toldo, Pietro Zanuttigh; Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7026-7035
Deep networks allow to obtain outstanding results in semantic segmentation, however they need to be trained in a single shot with a large amount of data. Continual learning settings where new classes are learned in incremental steps and previous training data is no longer available are challenging due to the catastrophic forgetting phenomenon. Existing approaches typically fail when several incremental steps are performed or in presence of a distribution shift of the background class. We tackle these issues by recreating no longer available data for the old classes and outlining a content inpainting scheme on the background class. We propose two sources for replay data. The first resorts to a generative adversarial network to sample from the class space of past learning steps. The second relies on web-crawled data to retrieve images containing examples of old classes from online databases. In both scenarios no samples of past steps are stored, thus avoiding privacy concerns. Replay data are then blended with new samples during the incremental steps. Our approach, RECALL, outperforms state-of-the-art methods.
****

Extending Neural P-Frame Codecs for B-Frame Coding
Reza Pourreza, Taco Cohen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6680-6689
While most neural video codecs address P-frame coding (predicting each frame from past ones), in this paper we address B-frame compression (predicting frames using both past and future reference frames). Our B-frame solution is based on the existing P-frame methods. As a result, B-frame coding capability can easily be added to an existing neural codec. The basic idea of our B-frame coding method is to interpolate the two reference frames to generate a single reference frame and then use it together with an existing P-frame codec to encode the input B-frame. Our studies show that the interpolated frame is a much better reference for the P-frame codec compared to using the previous frame as is usually done. Our results show that using the proposed method with an existing P-frame codec can lead to 28.5% saving in bit-rate on the UVG dataset compared to the P-frame codec while generating the same video quality.
****

HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering
Fei Liu, Jing Liu, Weining Wang, Hanqing Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1698-1707
Relational reasoning is at the heart of video question answering. However, existing approaches suffer from several common limitations: (1) they only focus on either object-level or frame-level relational reasoning, and fail to integrate the both; and (2) they neglect to leverage semantic knowledge for relational reasoning. In this work, we propose a Hierarchical VisuAl-Semantic RelatIonal Reasoning (HAIR) framework to address these limitations. Specifically, we present a novel graph memory mechanism to perform relational reasoning, and further develop two types of graph memory: a) visual graph memory that leverages visual information of video for relational reasoning; b) semantic graph memory that is specifically designed to explicitly leverage semantic knowledge contained in the classes and attributes of video objects, and perform relational reasoning in the semantic space. Taking advantage of both graph memory mechanisms, we build a hierarchical framework to enable visual-semantic relational reasoning from object level to frame level. Experiments on four challenging benchmark datasets show that the proposed framework leads to state-of-the-art performance, with fewer parameters and faster inference speed. Besides, our approach also shows superior performance on other video+language task.
****

Ensemble Attention Distillation for Privacy-Preserving Federated Learning
Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, Arun Innanje; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15076-15086

We consider the problem of Federated Learning (FL) where numerous decentralized computational nodes collaborate with each other to train a centralized machine learning model without explicitly sharing their local data samples. Such decentralized training naturally leads to issues of imbalanced or differing data distributions among the local models and challenges in fusing them into a central model. Existing FL methods deal with these issues by either sharing local parameters or fusing models via online distillation. However, such a design leads to multiple rounds of inter-node communication resulting in substantial bandwidth consumption, while also increasing the risk of data leakage and consequent privacy issues. To address these problems, we propose a new distillation-based FL framework that can preserve privacy by design, while also consuming substantially less network communication resources when compared to the current state-of-the-art. Our framework engages in inter-node communication using only publicly available and approved datasets, thereby giving explicit privacy control to the user. To distill knowledge among the various local models, our framework involves a novel ensemble distillation algorithm that uses both final prediction as well as model attention. This algorithm explicitly considers the diversity among various local nodes while also seeking consensus among them. This results in a comprehensive technique to distill knowledge from various decentralized nodes. We demonstrate the various aspects and the associated benefits of our FL framework through extensive experiments that produce state-of-the-art results on both classification and segmentation tasks on natural and medical images.
****

Voxel Transformer for 3D Object Detection
Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, Chunjing Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3164-3173
We present Voxel Transformer (VoTr), a novel and effective voxel-based Transformer backbone for 3D object detection from point clouds. Conventional 3D convolutional backbones in voxel-based 3D detectors cannot efficiently capture large context information, which is crucial for object recognition and localization, owing to the limited receptive fields. In this paper, we resolve the problem by introducing a Transformer-based architecture that enables long-range relationships between voxels by self-attention. Given the fact that non-empty voxels are naturally sparse but numerous, directly applying standard Transformer on voxels is non-trivial. To this end, we propose the sparse voxel module and the submanifold voxel module, which can operate on the empty and non-empty voxel positions effectively. To further enlarge the attention range while maintaining comparable computational overhead to the convolutional counterparts, we propose two attention mechanisms for multi-head attention in those two modules: Local Attention and Dilated Attention, and we further propose Fast Voxel Query to accelerate the querying process in multi-head attention. VoTr contains a series of sparse and submanifold voxel modules and can be applied in most voxel-based detectors. Our proposed VoTr shows consistent improvement over the convolutional baselines while maintaining computational efficiency on the KITTI dataset and the Waymo Open dataset.
****

Out-of-Boundary View Synthesis Towards Full-Frame Video Stabilization
Yufei Xu, Jing Zhang, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4842-4851
Warping-based video stabilizers smooth camera trajectory by constraining each pixel's displacement and warp stabilized frames from unstable ones accordingly. However, since the view outside the boundary is not available during warping, the resulting holes around the boundary of the stabilized frame must be discarded (i.e., cropping) to maintain visual consistency, and thus does leads to a tradeoff between stability and cropping ratio. In this paper, we make a first attempt to address this issue by proposing a new Out-of-boundary View Synthesis (OVS) method. By the nature of spatial coherence between adjacent frames and within each frame, OVS extrapolates the out-of-boundary view by aligning adjacent frames to each reference one. Technically, it first calculates the optical flow and propagates it to the outer boundary region according to the affinity, and then warps pi

xels accordingly. OVS can be integrated into existing warping-based stabilizers as a plug-and-play pre-processing module to significantly improve the cropping ratio of the stabilized results. In addition, stability is improved because the jitter amplification effect caused by cropping and resizing is reduced. Experimental results on the NUS benchmark show that OVS can improve the performance of five representative state-of-the-art methods in terms of objective metrics and subjective visual quality.
****

Multimodal Clustering Networks for Self-Supervised Learning From Unlabeled Videos
Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, Michael Picheny, Shih-Fu Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8012-8021
Multimodal self-supervised learning is getting more and more attention as it allows not only to train large networks without human supervision but also to search and retrieve data across various modalities. In this context, this paper proposes a framework that, starting from a pre-trained backbone, learns a common multimodal embedding space that, in addition to sharing representations across different modalities, enforces a grouping of semantically similar instances. To this end, we extend the concept of instance-level contrastive learning with a multimodal clustering step in the training pipeline to capture semantic similarities across modalities. The resulting embedding space enables retrieval of samples across all modalities, even from unseen datasets and different domains. To evaluate our approach, we train our model on the HowTo100M dataset and evaluate its zero-shot retrieval capabilities in two challenging domains, namely text-to-video retrieval, and temporal action localization, showing state-of-the-art results on four different datasets.
****

Towards a Universal Model for Cross-Dataset Crowd Counting
Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, Yihong Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3205-3214
This paper proposes to handle the practical problem of learning a universal model for crowd counting across scenes and datasets. We dissect that the crux of this problem is the catastrophic sensitivity of crowd counters to scale shift, which is very common in the real world and caused by factors such as different scene layouts and image resolutions. Therefore it is difficult to train a universal model that can be applied to various scenes. To address this problem, we propose scale alignment as a prime module for establishing a novel crowd counting framework. We derive a closed-form solution to get the optimal image rescaling factors for alignment by minimizing the distances between their scale distributions. A novel neural network together with a loss function based on an efficient sliced Wasserstein distance is also proposed for scale distribution estimation. Benefiting from the proposed method, we have learned a universal model that generally works well on several datasets where can even outperform state-of-the-art models that are particularly fine-tuned for each dataset significantly. Experiments also demonstrate the much better generalizability of our model to unseen scenes.
****

GAN Inversion for Out-of-Range Images With Geometric Transformations
Kyoungkook Kang, Seongtae Kim, Sunghyun Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13941-13949
For successful semantic editing of real images, it is critical for a GAN inversion method to find an in-domain latent code that aligns with the domain of a pre-trained GAN model. Unfortunately, such in-domain latent codes can be found only for in-range images that align with the training images of a GAN model. In this paper, we propose BDInvert, a novel GAN inversion approach to semantic editing of out-of-range images that are geometrically unaligned with the training images of a GAN model. To find a latent code that is semantically editable, BDInvert inverts an input out-of-range image into an alternative latent space than the orig

inal latent space. We also propose a regularized inversion method to find a solu
tion that supports semantic editing in the alternative space. Our experiments sh
ow that BDInvert effectively supports semantic editing of out-of-range images wi
th geometric transformations.
****

Scaling Up Instance Annotation via Label Propagation
Dim P. Papadopoulos, Ethan Weber, Antonio Torralba; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2021, pp. 15364-15373
Manually annotating object segmentation masks is very time-consuming. While inte
ractive segmentation methods offer a more efficient alternative, they become una
ffordable at a large scale because the cost grows linearly with the number of an
notated masks. In this paper, we propose a highly efficient annotation scheme fo
r building large datasets with object segmentation masks. At a large scale, imag
es contain many object instances with similar appearance. We exploit these simil
arities by using hierarchical clustering on mask predictions made by a segmentat
ion model. We propose a scheme that efficiently searches through the hierarchy o
f clusters and selects which clusters to annotate. Humans manually verify only a
 few masks per cluster, and the labels are propagated to the whole cluster. Thro
ugh a large-scale experiment to populate 1M unlabeled images with object segment
ation masks for 80 object classes, we show that (1) we obtain 1M object segmenta
tion masks with an total annotation time of only 290 hours; (2) we reduce annota
tion time by 76x compared to manual annotation; (3) the segmentation quality of
our masks is on par with those from manually annotated datasets. Code, data, and
 models are available online.
****

Learning RAW-to-sRGB Mappings With Inaccurately Aligned Supervision
Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, Wangmeng Zuo; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
1, pp. 4348-4358
Learning RAW-to-sRGB mapping has drawn increasing attention in recent years, whe
rein an input raw image is trained to imitate the target sRGB image captured by
another camera. However, the severe color inconsistency makes it very challengin
g to generate well-aligned training pairs of input raw and target sRGB images. W
hile learning with inaccurately aligned supervision is prone to causing pixel sh
ift and producing blurry results. In this paper, we circumvent such issue by pre
senting a joint learning model for image alignment and RAW-to-sRGB mapping. To d
iminish the effect of color inconsistency in image alignment, we introduce to us
e a global color mapping (GCM) module to generate an initial sRGB image given th
e input raw image, which can keep the spatial location of the pixels unchanged,
and the target sRGB image is utilized to guide GCM for converting the color towa
rds it. Then a pre-trained optical flow estimation network (e.g., PWC-Net) is de
ployed to warp the target sRGB image to align with the GCM output. To alleviate
the effect of inaccurately aligned supervision, the warped target sRGB image is
leveraged to learn RAW-to-sRGB mapping. When training is done, the GCM module an
d optical flow network can be detached, thereby bringing no extra computation co
st for inference. Experiments show that our method performs favorably against st
ate-of-the-arts on ZRR and SR-RAW datasets. With our joint learning model, a lig
ht-weight backbone can achieve better quantitative and qualitative performance o
n ZRR dataset. Codes are available at https://github.com/cszhilu1998/RAW-to-sRGB
.
****

Context Reasoning Attention Network for Image Super-Resolution
Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, Yun Fu; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 4278-4287
Deep convolutional neural networks (CNNs) are achieving great successes for imag
e super-resolution (SR), where global context is crucial for accurate restoratio
n. However, the basic convolutional layer in CNNs is designed to extract local p
atterns, lacking the ability to model global context. Many efforts have been dev
oted to augmenting SR networks with the global context information, especially b

y global feature interaction methods. These works incorporate the global context into local feature representation. However, recent advances in neuroscience show that it is necessary for the neurons to dynamically modulate their functions according to context, which is neglected in most CNN based SR methods. Motivated by those observations and analyses, we propose context reasoning attention network (CRAN) to adaptively modulate the convolution kernel according to the global context. Specifically, we extract global context descriptors, which are further enhanced with semantic reasoning. Channel and spatial interactions are then proposed to generate context reasoning attention mask, which is applied to modify the convolution kernel adaptively. Such a modulated convolution layer is utilized as basic component to build the network blocks and itself. Extensive experiments on benchmark datasets with multiple degradation models show that our CRAN achieves superior SR results and favourable efficiency trade-off.
****

FastNeRF: High-Fidelity Neural Rendering at 200FPS
Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, Julien Valentin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14346-14355
Recent work on Neural Radiance Fields (NeRF) showed how neural networks can be used to encode complex 3D environments that can be rendered photorealistically from novel viewpoints. Rendering these images is very computationally demanding and recent improvements are still a long way from enabling interactive rates, even on high-end hardware. Motivated by scenarios on mobile and mixed reality devices, we propose FastNeRF, the first NeRF-based system capable of rendering high fidelity photorealistic images at 200Hz on a high-end consumer GPU. The core of our method is a graphics-inspired factorization that allows for (i) compactly caching a deep radiance map at each position in space, (ii) efficiently querying that map using ray directions to estimate the pixel values in the rendered image. Extensive experiments show that the proposed method is 3000 times faster than the original NeRF algorithm and at least an order of magnitude faster than existing work on accelerating NeRF, while maintaining visual quality and extensibility.
****

3DeepCT: Learning Volumetric Scattering Tomography of Clouds
Yael Sde-Chen, Yoav Y. Schechner, Vadim Holodovsky, Eshkol Eytan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5671-5682
We present 3DeepCT, a deep neural network for computed tomography, which performs 3D reconstruction of scattering volumes from multi-view images. The architecture is dictated by the stationary nature of atmospheric cloud fields. The task of volumetric scattering tomography aims at recovering a volume from its 2D projections. This problem has been approached by diverse inverse methods based on signal processing and physics models. However, such techniques are typically iterative, exhibiting a high computational load and a long convergence time. We show that 3DeepCT outperforms physics-based inverse scattering methods, in accuracy, as well as offering orders of magnitude improvement in computational run-time. We further introduce a hybrid model that combines 3DeepCT and physics-based analysis. The resultant hybrid technique enjoys fast inference time and improved recovery performance.
****

EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-Resolution
Jin Han, Yixin Yang, Chu Zhou, Chao Xu, Boxin Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4882-4891
An event camera detects the scene radiance changes and sends a sequence of asynchronous event streams with high dynamic range, high temporal resolution, and low latency. However, the spatial resolution of event cameras is limited as a trade-off for these outstanding properties. To reconstruct high-resolution intensity images from event data, we propose EvIntSR-Net that converts event data to multiple latent intensity frames to achieve super-resolution on intensity images in this paper. EvIntSR-Net bridges the domain gap between event streams and intensit

y frames and learns to merge a sequence of latent intensity frames in a recurren t updating manner. Experimental results show that EvIntSR-Net can reconstruct SR intensity images with higher dynamic range and fewer blurry artifacts by fusing events with intensity frames for both simulated and real-world data. Furthermor e, the proposed EvIntSR-Net is able to generate high-frame-rate videos with supe r-resolved frames.

****

Deep Structured Instance Graph for Distilling Object Detectors

Yixin Chen, Pengguang Chen, Shu Liu, Liwei Wang, Jiaya Jia; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4359-4368

Effectively structuring deep knowledge plays a pivotal role in transfer from tea cher to student, especially in semantic vision tasks. In this paper, we present a simple knowledge structure to exploit and encode information inside the detect ion system to facilitate detector knowledge distillation. Specifically, aiming a t solving the feature imbalance problem while further excavating the missing rel ation inside semantic instances, we design a graph whose nodes correspond to ins tance proposal-level features and edges represent the relation between nodes. To further refine this graph, we design an adaptive background loss weight to redu ce node noise and background samples mining to prune trivial edges. We transfer the entire graph as encoded knowledge representation from teacher to student, ca pturing local and global information simultaneously. We achieve new state-of-the -art results on the challenging COCO object detection task with diverse student- teacher pairs on both one- and two-stage detectors. We also experiment with inst ance segmentation to demonstrate robustness of our method. It is notable that di stilled Faster R-CNN with ResNet18-FPN and ResNet50-FPN yields 38.68 and 41.82 B ox AP respectively on the COCO benchmark, Faster R-CNN with ResNet101-FPN signif icantly achieves 43.38 AP, which outperforms ResNet152-FPN teacher about 0.7 AP. Code: https://github.com/dvlab-research/Dsig.

****

Adversarial Attacks Are Reversible With Natural Supervision

Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, Carl Vondrick; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 661 -671

We find that images contain intrinsic structure that enables the reversal of man y adversarial attacks. Attack vectors cause not only image classifiers to fail, but also collaterally disrupt incidental structure in the image. We demonstrate that modifying the attacked image to restore the natural structure will reverse many types of attacks, providing a defense. Experiments demonstrate significantl y improved robustness for several state-of-the-art models across the CIFAR-10, C IFAR-100, SVHN, and ImageNet datasets. Our results show that our defense is stil l effective even if the attacker is aware of the defense mechanism. Since our de fense is deployed during inference instead of training, it is compatible with pr e-trained networks as well as most other defenses. Our results suggest deep netw orks are vulnerable to adversarial examples partly because their representations do not enforce the natural structure of images.

****

Hierarchical Graph Attention Network for Few-Shot Visual-Semantic Learning

Chengxiang Yin, Kun Wu, Zhengping Che, Bo Jiang, Zhiyuan Xu, Jian Tang; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 2177-2186

Deep learning has made tremendous success in computer vision, natural language p rocessing and even visual-semantic learning, which requires a huge amount of lab eled training data. Nevertheless, the goal of human-level intelligence is to ena ble a model to quickly obtain an in-depth understanding given a small number of samples, especially with heterogeneity in the multi-modal scenarios such as visu al question answering and image captioning. In this paper, we study the few-shot visual-semantic learning and present the Hierarchical Graph ATtention network (HGAT). This two-stage network models the intra- and inter-modal relationships wi th limited image-text samples. The main contributions of HGAT can be summarized as follows: 1) it sheds light on tackling few-shot multi-modal learning problems

, which focuses primarily, but not exclusively on visual and semantic modalities , through better exploitation of the intra-relationship of each modality and an attention-based co-learning framework between modalities using a hierarchical gr aph-based architecture; 2) it achieves superior performance on both visual quest ion answering and image captioning in the few-shot setting; 3) it can be easily extended to the semi-supervised setting where image-text samples are partially u nlabeled. We show via extensive experiments that HGAT delivers state-of-the-art performance on three widely-used benchmarks of two visual-semantic learning task s.

****

Semantics Disentangling for Generalized Zero-Shot Learning

Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, Zheng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8712-8720

Generalized zero-shot learning (GZSL) aims to classify samples under the assumpt ion that some classes are not observable during training. To bridge the gap betw een the seen and unseen classes, most GZSL methods attempt to associate the visu al features of seen classes with attributes or to generate unseen samples direct ly. Nevertheless, the visual features used in prior approaches do not necessaril y encode semantically related information that the shared attributes refer to, w hich greatly degrades the model generalization to unseen classes. To address thi s issue, in this paper, we propose a novel semantics disentangling framework for the generalized zero-shot learning task (SDGZSL), where the visual features dep icted unseen classes are firstly estimated by a conditional VAE and then factori zed into semantic-consistent and semantic-unrelated latent vectors. In particula r, a total correlation penalty is applied to guarantee the independence between the two factorized representations, and the semantic consistency of which is mea sured by the derived relation network. Extensive experiments conducted on four G ZSL benchmark datasets have evidenced that the semantic-consistent features dise ntangled by the proposed SDGZSL are more generalizable in tasks of canonical and generalized zero-shot learning.

****

Space-Time-Separable Graph Convolutional Network for Pose Forecasting

Theodoros Sofianos, Alessio Sampieri, Luca Franco, Fabio Galasso; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1120 9-11218

Human pose forecasting is a complex structured-data sequence-modelling task, whi ch has received increasing attention, also due to numerous potential application s. Research has mainly addressed the temporal dimension as time series and the i nteraction of human body joints with a kinematic tree or by a graph. This has de coupled the two aspects and leveraged progress from the relevant fields, but it has also limited the understanding of the complex structural joint spatio-tempor al dynamics of the human pose. Here we propose a novel Space-Time-Separable Grap h Convolutional Network (STS-GCN) for pose forecasting. For the first time, STS- GCN models the human pose dynamics only with a graph convolutional network (GCN) , including the temporal evolution and the spatial joint interaction within a si ngle-graph framework, which allows the cross-talk of motion and spatial correlat ions. Concurrently, STS-GCN is the first space-time-separable GCN: the space-tim e graph connectivity is factored into space and time affinity matrices, which bo ttlenecks the space-time cross-talk, while enabling full joint-joint and time-ti me correlations. Both affinity matrices are learnt end-to-end, which results in connections substantially deviating from the standard kinematic tree and the lin ear-time time series. In experimental evaluation on three complex, recent and la rge-scale benchmarks, Human3.6M [Ionescu et al. TPAMI'14], AMASS [Mahmood et al. ICCV'19] and 3DPW [Von Marcard et al. ECCV'18], STS-GCN outperforms the state-o f-the-art, surpassing the current best technique [Mao et al. ECCV'20] by over 32 % in average at the most difficult long-term predictions, while only requiring 1 .7% of its parameters. We explain the results qualitatively and illustrate the g raph interactions by the factored joint-joint and time-time learnt graph connect ions. Our source code is available at https://github.com/FraLuca/STSGCN

****
3D-FRONT: 3D Furnished Rooms With layOuts and semaNTics
Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, Hao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10933-10942
We introduce 3D-FRONT (3D Furnished Rooms with layOuts and semaNTics), a new, large-scale, and compre- hensive repository of synthetic indoor scenes highlighted by professionally designed layouts and a large number of rooms populated by high-quality textured 3D models with style compatibility. From layout semantics down to texture details of individual objects, our dataset is freely available to the academic community and beyond. Currently, 3D- FRONT contains 6,813 CAD houses, where 18,968 rooms diversely furnished by 3D objects, far surpassing all publicly available scene datasets. The 13,151 furniture objects all come with high-quality textures. While the floorplans and layout designs (i.e., furniture arrangements) are directly sourced from professional creations, the interior de- signs in terms of furniture styles, color, and textures have been carefully curated based on a recommender system we develop to attain consistent styles as expert designs. Furthermore, we release Trescope, a light-weight rendering tool, to support benchmark rendering of 2D images and annotations from 3D-FRONT. We demonstrate two applications, interior scene synthesis and texture synthesis, that are especially tailored to the strengths of our new dataset.
****
Meta-Learning With Task-Adaptive Loss Function for Few-Shot Learning
Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9465-9474
In few-shot learning scenarios, the challenge is to generalize and perform well on new unseen examples when only very few labeled examples are available for each task. Model-agnostic meta-learning (MAML) has gained the popularity as one of the representative few-shot learning methods for its flexibility and applicability to diverse problems. However, MAML and its variants often resort to a simple loss function without any auxiliary loss function or regularization terms that can help achieve better generalization. The problem lies in that each application and task may require different auxiliary loss function, especially when tasks are diverse and distinct. Instead of attempting to hand-design an auxiliary loss function for each application and task, we introduce a new meta-learning framework with a loss function that adapts to each task. Our proposed framework, named Meta-Learning with Task-Adaptive Loss Function (MeTAL), demonstrates the effectiveness and the flexibility across various domains, such as few-shot classification and few-shot regression.
****
Learning To Track Objects From Unlabeled Videos
Jilai Zheng, Chao Ma, Houwen Peng, Xiaokang Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13546-13555
In this paper, we propose to learn an Unsupervised Single Object Tracker (USOT) from scratch. We identify that three major challenges, i.e., moving object discovery, rich temporal variation exploitation, and online update, are the central causes of the performance bottleneck of existing unsupervised trackers. To narrow the gap between unsupervised trackers and supervised counterparts, we propose an effective unsupervised learning approach composed of three stages. First, we sample sequentially moving objects with unsupervised optical flow and dynamic programming, instead of random cropping. Second, we train a naive Siamese tracker from scratch using single-frame pairs. Third, we continue training the tracker with a novel cycle memory learning scheme, which is conducted in longer temporal spans and also enables our tracker to update online. Extensive experiments show that the proposed USOT learned from unlabeled videos performs well over the state-of-the-art unsupervised trackers by large margins, and on par with recent supervised deep trackers. Code is available at https://github.com/VISION-SJTU/USOT.
****
(Just) A Spoonful of Refinements Helps the Registration Error Go Down

Sérgio Agostinho, Aljoša Ošep, Alessio Del Bue, Laura Leal-Taixé; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6108-6117

In this paper, we tackle data-driven 3D point cloud registration. Given point correspondences, the standard Kabsch algorithm provides an optimal rotation estimate. This allows to train registration models in an end-to-end manner by differentiating the SVD operation. However, given the initial rotation estimate supplied by Kabsch, we show we can improve point correspondence learning during model training by extending the original optimization problem. In particular, we linearize the governing constraints of the rotation matrix and solve the resulting linear system of equations. We then iteratively produce new solutions by updating the initial estimate. Our experiments show that, by plugging our differentiable layer to existing learning-based registration methods, we improve the correspondence matching quality. This yields up to a 7% decrease in rotation error for correspondence-based data-driven registration methods.

****

H2O: A Benchmark for Visual Human-Human Object Handover Analysis

Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15762-15771

Object handover is a common human collaboration behavior that attracts attention from researchers in Robotics and Cognitive Science. Though visual perception plays an important role in the object handover task, the whole handover process has been specifically explored. In this work, we propose a novel rich-annotated dataset, H2O, for visual analysis of human-human object handovers. The H2O, which contains 18K video clips involving 15 people who hand over 30 objects to each other, is a multi-purpose benchmark. It can support several vision-based tasks, from which, we specifically provide a baseline method, RGPNet, for a less-explored task named Receiver Grasp Prediction. Extensive experiments show that the RGPNet can produce plausible grasps based on the giver's hand-object states in the pre-handover phase. Besides, we also report the hand and object pose errors with existing baselines and show that the dataset can serve as the video demonstrations for robot imitation learning on the handover task.

****

ECS-Net: Improving Weakly Supervised Semantic Segmentation by Using Connections Between Class Activation Maps

Kunyang Sun, Haoqing Shi, Zhengming Zhang, Yongming Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7283-7292

Image-level weakly supervised semantic segmentation is a challenging task. As classification networks tend to capture notable object features and are insensitive to overactivation, class activation map (CAM) is too sparse and rough to guide segmentation network training. Inspired by the fact that erasing distinguishing features force networks to collect new ones from non-discriminative object regions, we using relationships between CAMs to propose a novel weakly supervised method. In this work, we apply these features, learned from erased images, as segmentation supervision, driving network to study robust representation. In specifically, object regions obtained by CAM techniques are erased on images firstly. To provide other regions with segmentation supervision, Erased CAM Supervision Net (ECSNet) generates pixel-level labels by predicting segmentation results of those processed images. We also design the rule of suppressing noise to select reliable labels. Our experiments on PASCAL VOC 2012 dataset show that without data annotations except for ground truth image-level labels, our ECS-Net achieves 67.6% mIoU on test set and 66.6% mIoU on val set, outperforming previous state-of-the-art methods.

****

Heterogeneous Relational Complement for Vehicle Re-Identification

Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 205-214

The crucial problem in vehicle re-identification is to find the same vehicle identity when reviewing this object from cross-view cameras, which sets a higher de

mand for learning viewpoint-invariant representations. In this paper, we propose to solve this problem from two aspects: constructing robust feature representations and proposing camera-sensitive evaluations. We first propose a novel Heterogeneous Relational Complement Network (HRCN) by incorporating region-specific features and cross-level features as complements for the original high-level output. Considering the distributional differences and semantic misalignment, we propose graph-based relation modules to embed these heterogeneous features into one unified high-dimensional space. On the other hand, considering the deficiencies of cross-camera evaluations in existing measures (i.e., CMC and AP), we then propose a Cross-camera Generalization Measure (CGM) to improve the evaluations by introducing position-sensitivity and cross-camera generalization penalties. We further construct a new benchmark of existing models with our proposed CGM and experimental results reveal that our proposed HRCN model achieves new state-of-the-art in VeRi-776, VehicleID, and VERI-Wild.

****

Hierarchical Object-to-Zone Graph for Object Navigation

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, Shuqiang Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15130-15140

The goal of object navigation is to reach the expected objects according to visual information in the unseen environments. Previous works usually implement deep models to train an agent to predict actions in real-time. However, in the unseen environment, when the target object is not in egocentric view, the agent may not be able to make wise decisions due to the lack of guidance. In this paper, we propose a hierarchical object-to-zone (HOZ) graph to guide the agent in a coarse-to-fine manner, and an online-learning mechanism is also proposed to update HOZ according to the real-time observation in new environments. In particular, the HOZ graph is composed of scene nodes, zone nodes and object nodes. With the pre-learned HOZ graph, the real-time observation and the target goal, the agent can constantly plan an optimal path from zone to zone. In the estimated path, the next potential zone is regarded as sub-goal, which is also fed into the deep reinforcement learning model for action prediction. Our methods are evaluated on the AI2-Thor simulator. In addition to widely used evaluation metrics SR and SPL, we also propose a new evaluation metric of SAE that focuses on the effective action rate. Experimental results demonstrate the effectiveness and efficiency of our proposed method. The code is available at https://github.com/sx-zhang/HOZ.git.

****

Information-Theoretic Regularization for Multi-Source Domain Adaptation

Geon Yeong Park, Sang Wan Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9214-9223

Adversarial learning strategy has demonstrated remarkable performance in dealing with single-source Domain Adaptation (DA) problems, and it has recently been applied to Multi-source DA (MDA) problems. Although most existing MDA strategies rely on a multiple domain discriminator setting, its effect on the latent space representations has been poorly understood. Here we adopt an information-theoretic approach to identify and resolve the potential adverse effect of the multiple domain discriminators on MDA: disintegration of domain-discriminative information, limited computational scalability, and a large variance in the gradient of the loss during training. We examine the above issues by situating adversarial DA in the context of information regularization. This also provides a theoretical justification for using a single and unified domain discriminator. Based on this idea, we implement a novel neural architecture called a Multi-source Information-regularized Adaptation Networks (MIAN). Large-scale experiments demonstrate that MIAN, despite its structural simplicity, reliably and significantly outperforms other state-of-the-art methods.

****

Beyond Road Extraction: A Dataset for Map Update Using Aerial Images

Favyen Bastani, Samuel Madden; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11905-11914

The increasing availability of satellite and aerial imagery has sparked substant

ial interest in automatically updating street maps by processing aerial images. Until now, the community has largely focused on road extraction, where road networks are inferred from scratch from an aerial image. However, given that relatively high-quality maps exist in most parts of the world, in practice, inference approaches must be applied to update existing maps rather than infer new ones. With recent road extraction methods showing high accuracy, we argue that it is time to transition to the more practical map update task, where an existing map is updated by adding, removing, and shifting roads, without introducing errors in parts of the existing map that remain up-to-date. In this paper, we develop a new dataset called MUNO21 for the map update task, and show that it poses several new and interesting research challenges. We evaluate several state-of-the-art road extraction methods on MUNO21, and find that substantial further improvements in accuracy will be needed to realize automatic map update.

****

Transfusion: A Novel SLAM Method Focused on Transparent Objects

Yifan Zhu, Jiaxiong Qiu, Bo Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6019-6028

Recently RGB-D sensors have become very popular in the area of Simultaneous Localisation and Mapping (SLAM). The RGB-D SLAM approach relies heavily on the accuracy of the input depth map. However, refraction and reflection of transparent objects will result in false depth input of RGB-D cameras, which makes the traditional RGB-D SLAM algorithm unable to work correctly in the presence of transparent objects. In this paper, we propose a novel SLAM approach called transfusion that allows transparent object existence and recovery in the video input. Our method is composed of two parts. Transparent Objects Cut Iterative Closest Points (TC-ICP)is first used to recover camera pose, detecting and removing transparent objects from input to reduce the trajectory errors. Then Transparent Objects Reconstruction (TO-Reconstruction) is used to reconstruct the transparent objects and opaque objects separately. The opaque objects are reconstructed with the traditional method, and the transparent objects are reconstructed with the visual hull-based method. To evaluate our algorithm, we construct a new RGB-D SLAM database containing 25 video sequences. Each sequence has at least one transparent object. Experiments show that our approach can work adequately in scenes contain transparent objects while the existing approach can not handle them. Our approach significantly improves the accuracy of the camera trajectory and the quality of environment reconstruction.

****

Physics-Based Human Motion Estimation and Synthesis From Videos

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, Florian Shkurti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11532-11541

Human motion synthesis is an important problem for applications in graphics and gaming, and even in simulation environments for robotics. Existing methods require accurate motion capture data for training, which is costly to obtain. Instead, we propose a framework for training generative models of physically plausible human motion directly from monocular RGB videos, which are much more widely available. At the core of our method is a novel optimization formulation that aims to correct imperfect image-based pose estimations by enforcing physics constraints and reasons about contacts in a differentiable way. This optimization yields corrected 3D poses and motions, as well as their corresponding contact forces. Results show that our physically-correct motions significantly outperform prior work on pose estimation. We then train a generative model to synthesize both future motion and contact forces. We demonstrate both qualitatively and quantitatively significantly improved motion synthesis quality and physical plausibility achieved by our method on the large scale Human3.6m dataset as compared to prior learning-based kinematic and physics-based methods. By learning directly from video, our method paves the way for large-scale, realistic and diverse motion synthesis not previously possible.

****

Hierarchical Memory Matching Network for Video Object Segmentation

Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, Euntai Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12889-12898

We present Hierarchical Memory Matching Network (HMMN) for semi-supervised video object segmentation. Based on a recent memory-based method [33], we propose two advanced memory read modules that enable us to perform memory reading in multiple scales while exploiting temporal smoothness. We first propose a kernel guided memory matching module that replaces the non-local dense memory read, commonly adopted in previous memory-based methods. The module imposes the temporal smoothness constraint in the memory read, leading to accurate memory retrieval. More importantly, we introduce a hierarchical memory matching scheme and propose a top-k guided memory matching module in which memory read on a fine-scale is guided by that on a coarse-scale. With the module, we perform memory read in multiple scales efficiently and leverage both high-level semantic and low-level fine-grained memory features to predict detailed object masks. Our network achieves state-of-the-art performance on the validation sets of DAVIS 2016/2017 (90.8% and 84.7%) and YouTube-VOS 2018/2019 (82.6% and 82.5%), and test-dev set of DAVIS 2017 (78.6%). The source code and model are available online: https://github.com/Hongje/HMMN.

****

Pathdreamer: A World Model for Indoor Navigation

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, Peter Anderson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14738-14748

People navigating in unfamiliar buildings take advantage of myriad visual, spatial and semantic cues to efficiently achieve their navigation goals. Towards equipping computational agents with similar capabilities, we introduce Pathdreamer, a visual world model for agents navigating in novel indoor environments. Given one or more previous visual observations, Pathdreamer generates plausible high-resolution 360deg visual observations (RGB, semantic segmentation and depth) for viewpoints that have not been visited, in buildings not seen during training. In regions of high uncertainty (e.g. predicting around corners, imagining the contents of an unseen room), Pathdreamer can predict diverse scenes, allowing an agent to sample multiple realistic outcomes for a given trajectory. We demonstrate that Pathdreamer encodes useful and accessible visual, spatial and semantic knowledge about human environments by using it in the downstream task of Vision-and-Language Navigation (VLN). Specifically, we show that planning ahead with Pathdreamer brings about half the benefit of looking ahead at actual observations from unobserved parts of the environment. We hope that Pathdreamer will help unlock model-based approaches to challenging embodied navigation tasks such as navigating to specified objects and VLN.

****

Saliency-Associated Object Tracking

Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, Zhenyu He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9866-9875

Most existing trackers based on deep learning perform tracking in a holistic strategy, which aims to learn deep representations of the whole target for localizing the target. It is arduous for such methods to track targets with various appearance variations. To address this limitation, another type of methods adopts a part-based tracking strategy which divides the target into equal patches and tracks all these patches in parallel. The target state is inferred by summarizing the tracking results of these patches. A potential limitation of such trackers is that not all patches are equally informative for tracking. Some patches that are not discriminative may have adverse effects. In this paper, we propose to track the salient local parts of the target that are discriminative for tracking. In particular, we propose a fine-grained saliency mining module to capture the local saliencies. Further, we design a saliency-association modeling module to associate the captured saliencies together to learn effective correlation representations between the exemplar and the search image for state estimation. Extensive

experiments on five diverse datasets demonstrate that the proposed method perfor
ms favorably against state-of-the-art trackers.
****

Wanderlust: Online Continual Object Detection in the Real World
Jianren Wang, Xin Wang, Yue Shang-Guan, Abhinav Gupta; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2021, pp. 10829-10838
Online continual learning from data streams in dynamic environments is a critica
l direction in the computer vision field. However, realistic benchmarks and fund
amental studies in this line are still missing. To bridge the gap, we present a
new online continual object detection benchmark with an egocentric video dataset
, Objects Around Krishna (OAK). OAK adopts the KrishnaCAM videos, an ego-centric
 video stream collected over nine months by a graduate student. OAK provides exh
austive bounding box annotations of 80 video snippets (~17.5 hours) for 105 obje
ct categories in outdoor scenes. The emergence of new object categories in our b
enchmark follows a pattern similar to what a single person might see in their da
y-to-day life. The dataset also captures the natural distribution shifts as the
person travels to different places. These egocentric long running videos provide
 a realistic playground for continual learning algorithms, especially in online
embodied settings. We also introduce new evaluation metrics to evaluate the mode
l performance and catastrophic forgetting and provide baseline studies for onlin
e continual object detection. We believe this benchmark will pose new exciting c
hallenges for learning from non-stationary data in continual learning. The OAK d
ataset and the associated benchmark are released at https://oakdata.github.io/.
****

Distilling Optimal Neural Networks: Rapid Search in Diverse Spaces
Bert Moons, Parham Noorzad, Andrii Skliar, Giovanni Mariani, Dushyant Mehta, Chr
is Lott, Tijmen Blankevoort; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 12229-12238
Current state-of-the-art Neural Architecture Search (NAS) methods neither effici
ently scale to many hardware platforms nor handle diverse architectural search-s
paces. To remedy this, we present DONNA (Distilling Optimal Neural Network Archi
tectures), a novel pipeline for rapid, scalable and diverse NAS, that scales to
many user scenarios. DONNA consists of three phases. First, an accuracy predicto
r is built using blockwise knowledge distillation from a reference model. This p
redictor enables searching across diverse networks with varying macro-architectu
ral parameters such as layer types and attention mechanisms, as well as across m
icro-architectural parameters such as block repeats and expansion rates. Second,
 a rapid evolutionary search finds a set of pareto-optimal architectures for any
 scenario using the accuracy predictor and on-device measurements. Third, optima
l models are quickly finetuned to training-from-scratch accuracy. DONNA is up to
 100x faster than MNasNet in finding state-of-the-art architectures on-device. C
lassifying ImageNet, DONNA architectures are 20% faster than EfficientNet-B0 and
 MobileNetV2 on a Nvidia V100 GPU and 10% faster with 0.5% higher accuracy than
MobileNetV2-1.4x on a Samsung S20 smartphone. In addition to NAS, DONNA is used
for search-space extension and exploration, as well as hardware-aware model comp
ression.
****

Removing Adversarial Noise in Class Activation Feature Space
Dawei Zhou, Nannan Wang, Chunlei Peng, Xinbo Gao, Xiaoyu Wang, Jun Yu, Tongliang
 Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2021, pp. 7878-7887
Deep neural networks (DNNs) are vulnerable to adversarial noise. Pre-processing
based defenses could largely remove adversarial noise by processing inputs. Howe
ver, they are typically affected by the error amplification effect, especially i
n the front of continuously evolving attacks. To solve this problem, in this pap
er, we propose to remove adversarial noise by implementing a self-supervised adv
ersarial training mechanism in a class activation feature space. To be specific,
 we first maximize the disruptions to class activation features of natural examp
les to craft adversarial examples. Then, we train a denoising model to minimize
the distances between the adversarial examples and the natural examples in the c

lass activation feature space. Empirical evaluations demonstrate that our method could significantly enhance adversarial robustness in comparison to previous state-of-the-art approaches, especially against unseen adversarial attacks and adaptive attacks.

****

Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution

Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4096-4105

Existing blind image super-resolution (SR) methods mostly assume blur kernels are spatially invariant across the whole image. However, such an assumption is rarely applicable for real images whose blur kernels are usually spatially variant due to factors such as object motion and out-of-focus. Hence, existing blind SR methods would inevitably give rise to poor performance in real applications. To address this issue, this paper proposes a mutual affine network (MANet) for spatially variant kernel estimation. Specifically, MANet has two distinctive features. First, it has a moderate receptive field so as to keep the locality of degradation. Second, it involves a new mutual affine convolution (MAConv) layer that enhances feature expressiveness without increasing receptive field, model size and computation burden. This is made possible through exploiting channel interdependence, which applies each channel split with an affine transformation module whose input are the rest channel splits. Extensive experiments on synthetic and real images show that the proposed MANet not only performs favorably for both spatially variant and invariant kernel estimation, but also leads to state-of-the-art blind SR performance when combined with non-blind SR methods.

****

Unifying Nonlocal Blocks for Neural Networks

Lei Zhu, Qi She, Duo Li, Yanye Lu, Xuejing Kang, Jie Hu, Changhu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12292-12301

The nonlocal-based blocks are designed for capturing long-range spatial-temporal dependencies in computer vision tasks. Although having shown excellent performance, they still lack the mechanism to encode the rich, structured information among elements in an image or video. In this paper, to theoretically analyze the property of these nonlocal-based blocks, we provide a new perspective to interpret them, where we view them as a set of graph filters generated on a fully-connected graph. Specifically, when choosing the Chebyshev graph filter, a unified formulation can be derived for explaining and analyzing the existing nonlocal-based blocks (e.g., nonlocal block, nonlocal stage, double attention block). Furthermore, by concerning the property of spectral, we propose an efficient and robust spectral nonlocal block, which can be more robust and flexible to catch long-range dependencies when inserted into deep neural networks than the existing nonlocal blocks. Experimental results demonstrate the clear-cut improvements and practical applicabilities of our method on image classification, action recognition, semantic segmentation, and person re-identification tasks. Code are available at https://github.com/zh460045050/SNL_ICCV2021.

****

Learning Realistic Human Reposing Using Cyclic Self-Supervision With 3D Shape, Pose, and Appearance Consistency

Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry S. Davis, Javier Romero, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11138-11147

Synthesizing images of a person in novel poses from a single image is a highly ambiguous task. Most existing approaches require paired training images; i.e. images of the same person with the same clothing in different poses. However, obtaining sufficiently large datasets with paired data is challenging and costly. Previous methods that forego paired supervision lack realism. We propose a self-supervised framework named SPICE (Self-supervised Person Image CrEation) that closes the image quality gap with supervised methods. The key insight enabling self-s

upervision is to exploit 3D information about the human body in several ways. Fi rst, the 3D body shape must remain unchanged when reposing. Second, representing body pose in 3D enables reasoning about self occlusions. Third, 3D body parts t hat are visible before and after reposing, should have similar appearance featur es. Once trained, SPICE takes an image of a person and generates a new image of that person in a new target pose. SPICE achieves state-of-the-art performance on the DeepFashion dataset, improving the FID score from 29.9 to 7.8 compared with previous unsupervised methods, and with performance similar to the state-of-the -art supervised method (6.4). SPICE also generates temporally coherent videos gi ven an input image and a sequence of poses, despite being trained on static imag es only.

****

MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inve rse Reinforcement Learning

Sonia Baee, Erfan Pakdamanian, Inki Kim, Lu Feng, Vicente Ordonez, Laura Barnes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13178-13188

Inspired by human visual attention, we propose a novel inverse reinforcement lea rning formulation using Maximum Entropy Deep Inverse Reinforcement Learning (MED IRL) for predicting the visual attention of drivers in accident-prone situations . MEDIRL predicts fixation locations that lead to maximal rewards by learning a task-sensitive reward function from eye fixation patterns recorded from attentiv e drivers. Additionally, we introduce EyeCar, a new driver attention dataset in accident-prone situations. We conduct comprehensive experiments to evaluate our proposed model on three common benchmarks: (DR(eye)VE, BDD-A, DADA-2000), and ou r EyeCar dataset. Results indicate that MEDIRL outperforms existing models for p redicting attention and achieves state-of-the-art performance. We present extens ive ablation studies to provide more insights into different features of our pro posed model.

****

Simpler Is Better: Few-Shot Semantic Segmentation With Classifier Weight Transfo rmer

Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8741-8750

A few-shot semantic segmentation model is typically composed of a CNN encoder, a CNN decoder and a simple classifier (separating foreground and background pixel s). Most existing methods meta-learn all three model components for fast adaptat ion to a new class. However, given that as few as a single support set image is available, effective model adaption of all three components to the new class is extremely challenging. In this work we propose to simplify the meta-learning tas k by focusing solely on the simplest component -- the classifier, whilst leaving the encoder and decoder to pre-training. We hypothesize that if we pre-train an off-the-shelf segmentation model over a set of diverse training classes with su fficient annotations, the encoder and decoder can capture rich discriminative fe atures applicable for any unseen classes, rendering the subsequent meta-learning stage unnecessary. For the classifier meta-learning, we introduce a Classifier Weight Transformer (CWT) designed to dynamically adapt the support-set trained c lassifier's weights to each query image in an inductive way. Extensive experimen ts on two standard benchmarks show that despite its simplicity, our method outpe rforms the state-of-the-art alternatives, often by a large margin. Code is avail able on https://github.com/zhiheLu/CWT-for-FSS.

****

Prediction by Anticipation: An Action-Conditional Prediction Method Based on Int eraction Learning

Ershad Banijamali, Mohsen Rohani, Elmira Amirloo, Jun Luo, Pascal Poupart; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15621-15630

In autonomous driving (AD), accurately predicting changes in the environment can effectively improve safety and comfort. Due to complex interactions among traff

ic participants, however, it is very hard to achieve accurate prediction for a l
ong horizon. To address this challenge, we propose prediction by anticipation, w
hich views interaction in terms of a latent probabilistic generative process whe
rein some vehicles move partly in response to the anticipated motion of other ve
hicles. Under this view, consecutive data frames can be factorized into sequenti
al samples from an action-conditional distribution that effectively generalizes
to a wider range of actions and driving situations. Our proposed prediction mode
l, variational Bayesian in nature, is trained to maximize the evidence lower bou
nd (ELBO) of the log-likelihood of this conditional distribution. Evaluations of
 our approach with prominent AD datasets NGSIM I-80 and Argoverse show significa
nt improvement over current state-of-the-art in both accuracy and generalization
.
****

Scene Synthesis via Uncertainty-Driven Attribute Synchronization
Haitao Yang, Zaiwei Zhang, Siming Yan, Haibin Huang, Chongyang Ma, Yi Zheng, Cha
ndrajit Bajaj, Qixing Huang; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 5630-5640
Developing deep neural networks to generate 3D scenes is a fundamental problem i
n neural synthesis with immediate applications in architectural CAD, computer gr
aphics, as well as in generating virtual robot training environments. This task
is challenging because 3D scenes exhibit diverse patterns, ranging from continuo
us ones, such as object sizes and the relative poses between pairs of shapes, to
 discrete patterns, such as occurrence and co-occurrence of objects with symmetr
ical relationships. This paper introduces a novel neural scene synthesis approac
h that can capture diverse feature patterns of 3D scenes. Our method combines th
e strength of both neural network-based and conventional scene synthesis approac
hes. We use the parametric prior distributions learned from training data, which
 provide uncertainties of object attributes and relative attributes, to regulari
ze the outputs of feed-forward neural models. Moreover, instead of merely predic
ting a scene layout, our approach predicts an over-complete set of attributes. T
his methodology allows us to utilize the underlying consistency constraints amon
g the predicted attributes to prune infeasible predictions. Experimental results
 show that our approach outperforms existing methods considerably. The generated
 3D scenes interpolate the training data faithfully while preserving both contin
uous and discrete feature patterns.
****

Domain-Aware Universal Style Transfer
Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, Hyeran Byun; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14609-
14617
Style transfer aims to reproduce content images with the styles from reference i
mages. Existing universal style transfer methods successfully deliver arbitrary
styles to original images either in an artistic or a photo-realistic way. Howeve
r, the range of "arbitrary style" defined by existing works is bounded in the pa
rticular domain due to their structural limitation. Specifically, the degrees of
 content preservation and stylization are established according to a predefined
target domain. As a result, both photo-realistic and artistic models have diffic
ulty in performing the desired style transfer for the other domain. To overcome
this limitation, we propose a unified architecture, Domain-aware Style Transfer
Networks (DSTN) that transfer not only the style but also the property of domain
 (i.e., domainness) from a given reference image. To this end, we design a novel
 domainness indicator that captures the domainness value from the texture and st
ructural features of reference images. Moreover, we introduce a unified framewor
k with domainaware skip connection to adaptively transfer the stroke and palette
 to the input contents guided by the domainness indicator. Our extensive experim
ents validate that our model produces better qualitative results and outperforms
 previous methods in terms of proxy metrics on both artistic and photo-realistic
 stylizations. All codes and pre-trained weights are available at https://github
.com/Kibeom-Hong/Domain-Aware-Style-Transfer.
****

Adversarial Example Detection Using Latent Neighborhood Graph
Ahmed Abusnaina, Yuhang Wu, Sunpreet Arora, Yizhen Wang, Fei Wang, Hao Yang, Dav id Mohaisen; Proceedings of the IEEE/CVF International Conference on Computer Vi sion (ICCV), 2021, pp. 7687-7696
Detection of adversarial examples with high accuracy is critical for the securit y of deployed deep neural network-based models. We present the first graph-based adversarial detection method that constructs a Latent Neighborhood Graph (LNG) around an input example to determine if the input example is adversarial. Given an input example, selected reference adversarial and benign examples are used to capture the local manifold in the vicinity of the input example. The LNG node c onnectivity parameters are optimized jointly with the parameters of a graph atte ntion network in an end-to-end manner to determine the optimal graph topology fo r adversarial example detection. The graph attention network is used to determin e if the LNG is derived from an adversarial or benign input example. Experimenta l evaluations on CIFAR-10, STL-10, and ImageNet datasets, using six adversarial attack methods, demonstrate that the proposed method outperforms state-of-the-ar t adversarial detection methods in white-box and gray-box settings. The proposed method is able to successfully detect adversarial examples crafted with small p erturbations using unseen attacks.
****

Dynamic View Synthesis From Dynamic Monocular Video
Chen Gao, Ayush Saraf, Johannes Kopf, Jia-Bin Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5712-5721
We present an algorithm for generating novel views at arbitrary viewpoints and a ny input time step given a monocular video of a dynamic scene. Our work builds u pon recent advances in neural implicit representation and uses continuous and di fferentiable functions for modeling the time-varying structure and the appearanc e of the scene. We jointly train a time-invariant static NeRF and a time-varying dynamic NeRF, and learn how to blend the results in an unsupervised manner. How ever, learning this implicit function from a single video is highly ill-posed (w ith infinitely many solutions that match the input video). To resolve the ambigu ity, we introduce regularization losses to encourage a more physically plausible solution. We show extensive quantitative and qualitative results of dynamic vie w synthesis from casually captured videos.
****

Online Pseudo Label Generation by Hierarchical Cluster Dynamics for Adaptive Per son Re-Identification
Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglia n Qi, Dapeng Chen; Proceedings of the IEEE/CVF International Conference on Compu ter Vision (ICCV), 2021, pp. 8371-8381
Adaptive person re-identification (adaptive ReID) targets at transferring learne d knowledge from the labeled source domain to the unlabeled target domain. Pseud o-label-based methods that alternatively generate pseudo labels and optimize the training model have demonstrated great effectiveness in this field. However, th e generated pseudo labels are inaccurate and cannot reflect the true semantic me aning of the unlabeled samples. We consider such inaccuracy stems from both the lagged update of the pseudo labels as well as the simple criterion of the employ ed clustering method. To tackle the problem, we propose an online pseudo label g eneration by hierarchical cluster dynamics for adaptive ReID. In particular, hie rarchical label banks are constructed for all the samples in the dataset, and we update the pseudo labels of the sample in each coming mini-batch, performing th e model optimization and the label generation simultaneously. A new hierarchical cluster dynamics is built for the label update, where cluster merge and cluster split are driven by a possibility computed by the label propagation. Our method can achieve better pseudo labels and higher reid accuracy. Extensive experiment s on Market-to-Duke, Duke-to-Market, MSMT-to-Market, MSMT-to-Duke, Market-to-MSM T, and Duke-to-MSMT verify the effectiveness of our proposed method.
****

Learning To Match Features With Seeded Graph Matching Network
Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan

Tai, Long Quan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6301-6310
Matching local features across images is a fundamental problem in computer vision. Targeting towards high accuracy and efficiency, we propose Seeded Graph Matching Network, a graph neural network with sparse structure to reduce redundant connectivity and learn compact representation. The network consists of 1) Seeding Module, which initializes the matching by generating a small set of reliable matches as seeds. 2) Seeded Graph Neural Network, which utilizes seed matches to pass messages within/across images and predicts assignment costs. Three novel operations are proposed as basic elements for message passing: 1) Attentional Pooling, which aggregates keypoint features within the image to seed matches. 2) Seed Filtering, which enhances seed features and exchanges messages across images. 3) Attentional Unpooling, which propagates seed features back to original keypoints. Experiments show that our method reduces computational and memory complexity significantly compared with typical attention-based networks while competitive or higher performance is achieved.
****

Aha! Adaptive History-Driven Attack for Decision-Based Black-Box Models
Jie Li, Rongrong Ji, Peixian Chen, Baochang Zhang, Xiaopeng Hong, Ruixin Zhang, Shaoxin Li, Jilin Li, Feiyue Huang, Yongjian Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16168-16177
The decision-based black-box attack means to craft adversarial examples with only the top-1 label of the victim model available. A common practice is to start from a large perturbation and then iteratively reduce it with a deterministic direction and a random one while keeping it adversarial. The limited information obtained from each query and inefficient direction sampling impede attack efficiency, making it hard to obtain a small enough perturbation within a limited number of queries. To tackle this problem, we propose a novel attack method termed Adaptive History-driven Attack (AHA) which gathers information from all historical queries as the prior for current sampling. Moreover, to balance between the deterministic direction and the random one, we dynamically adjust the coefficient according to the ratio of the actual magnitude reduction to the expected one. Such a strategy improves the success rate of queries during optimization, letting adversarial examples move swiftly along the decision boundary. Our method can also integrate with subspace optimization like dimension reduction to further improve efficiency. Extensive experiments on both ImageNet and CelebA datasets demonstrate that our method achieves at least 24.3% lower magnitude of perturbation on average with the same number of queries. Finally, we prove the practical potential of our method by evaluating it on popular defense methods and a real-world system provided by MEGVII Face++.
****

Anonymizing Egocentric Videos
Daksh Thapar, Aditya Nigam, Chetan Arora; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2320-2329
In egocentric videos, the face of a wearer capturing the video is never captured. This gives a false sense of security that the wearer's privacy is preserved while sharing such videos. However, egocentric cameras are typically harnessed to wearer's head, and hence, also capture wearer's gait. Recent works have shown that wearer gait signatures can be extracted from egocentric videos, which can be used to determine if two egocentric videos have the same wearer. In a more damaging scenario, one can even recognize a wearer using hand gestures from egocentric videos, or identify a wearer in third person videos such as from a surveillance camera. We believe, this could be a death knell in sharing of egocentric videos, and fatal for egocentric vision research. In this work, we suggest a novel technique to anonymize egocentric videos, which create carefully crafted, but small, and imperceptible optical flow perturbations in an egocentric video's frames. Importantly, these perturbations do not affect object detection or action/activity recognition from egocentric videos but are strong enough to dis-balance the gait recovery process. In our experiments on benchmark \epic dataset, the proposed perturbation degrades the wearer recognition performance of [??], from 66.3%

to 13.4%, while preserving the activity recognition performance of [??] from 89.6% to 87.4%. To test our anonymization with more wearer recognition techniques, we also developed a stronger, and more generalizable wearer recognition method based on camera egomotion cues. The approach achieves state-of-the-art (SOTA) performance of 59.67% on \epicns, compared to 55.06% by [??]. However, the accuracy of our recognition technique also drops to 12% using the proposed anonymizing perturbations.

****

Modulated Graph Convolutional Network for 3D Human Pose Estimation

Zhiming Zou, Wei Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11477-11487

The graph convolutional network (GCN) has recently achieved promising performance of 3D human pose estimation (HPE) by modeling the relationship among body parts. However, most prior GCN approaches suffer from two main drawbacks. First, they share a feature transformation for each node within a graph convolution layer. This prevents them from learning different relations between different body joints. Second, the graph is usually defined according to the human skeleton and is suboptimal because human activities often exhibit motion patterns beyond the natural connections of body joints. To address these limitations, we introduce a novel Modulated GCN for 3D HPE. It consists of two main components: weight modulation and affinity modulation. Weight modulation learns different modulation vectors for different nodes so that the feature transformations of different nodes are disentangled while retaining a small model size. Affinity modulation adjusts the graph structure in a GCN so that it can model additional edges beyond the human skeleton. We investigate several affinity modulation methods as well as the impact of regularizations. Rigorous ablation study indicates both types of modulation improve performance with negligible overhead. Compared with state-of-the-art GCNs for 3D HPE, our approach either significantly reduces the estimation errors, e.g., by around 10%, while retaining a small model size or drastically reduces the model size, e.g., from 4.22M to 0.29M (a 14.5 X reduction), while achieving comparable performance. Results on two benchmarks show our Modulated GCN outperforms some recent states of the art. Our code is available at https://github.com/ZhimingZo/Modulated-GCN.

****

Learning Self-Consistency for Deepfake Detection

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, Wei Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15023-15033

We propose a new method to detect deepfake images using the cue of the source feature inconsistency within the forged images. It is based on the hypothesis that images' distinct source features can be preserved and extracted after going through state-of-the-art deepfake generation processes. We introduce a novel representation learning approach, called pair-wise self-consistency learning (PCL), for training ConvNets to extract these source features and detect deepfake images. It is accompanied by a new image synthesis approach, called inconsistency image generator (I2G), to provide richly annotated training data for PCL. Experimental results on seven popular datasets show that our models improve averaged AUC from 96.45% to 98.05% over the state of the art in the in-dataset evaluation and from 86.03% to 92.18% in the cross-dataset evaluation.

****

PreDet: Large-Scale Weakly Supervised Pre-Training for Detection

Vignesh Ramanathan, Rui Wang, Dhruv Mahajan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2865-2875

State-of-the-art object detection approaches typically rely on pre-trained classification models to achieve better performance and faster convergence. We hypothesize that classification pre-training strives to achieve translation invariance, and consequently ignores the localization aspect of the problem. We propose a new large-scale pre-training strategy for detection, where noisy class labels are available for all images, but not bounding-boxes. In this setting, we augment standard classification pre-training with a new detection-specific pretext task.

Motivated by the noise-contrastive learning based self-supervised approaches, we design a task that forces bounding boxes with high-overlap to have similar representations in different views of an image, compared to non-overlapping boxes. We redesign Faster R-CNN modules to perform this task efficiently. Our experimental results show significant improvements over existing weakly-supervised and self-supervised pre-training approaches in both detection accuracy as well as fine-tuning speed.

****

Real-Time Video Inference on Edge Devices via Adaptive Model Streaming

Mehrdad Khani, Pouya Hamadanian, Arash Nasr-Esfahany, Mohammad Alizadeh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4572-4582

Real-time video inference on edge devices like mobile phones and drones is challenging due to the high computation cost of Deep Neural Networks. We present Adaptive Model Streaming (AMS), a new approach to improving the performance of efficient lightweight models for video inference on edge devices. AMS uses a remote server to continually train and adapt a small model running on the edge device, boosting its performance on the live video using online knowledge distillation from a large, state-of-the-art model. We discuss the challenges of over-the-network model adaptation for video inference and present several techniques to reduce communication the cost of this approach: avoiding excessive overfitting, updating a small fraction of important model parameters, and adaptive sampling of training frames at edge devices. On the task of video semantic segmentation, our experimental results show 0.4--17.8 percent mean Intersection-over-Union improvement compared to a pre-trained model across several video datasets. Our prototype can perform video segmentation at 30 frames-per-second with 40 milliseconds camera-to-label latency on a Samsung Galaxy S10+ mobile phone, using less than 300 Kbps uplink and downlink bandwidth on the device.

****

Learning Generative Models of Textured 3D Meshes From Real-World Images

Dario Pavllo, Jonas Kohler, Thomas Hofmann, Aurelien Lucchi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13879-13889

Recent advances in differentiable rendering have sparked an interest in learning generative models of textured 3D meshes from image collections. These models natively disentangle pose and appearance, enable downstream applications in computer graphics, and improve the ability of generative models to understand the concept of image formation. Although there has been prior work on learning such models from collections of 2D images, these approaches require a delicate pose estimation step that exploits annotated keypoints, thereby restricting their applicability to a few specific datasets. In this work, we propose a GAN framework for generating textured triangle meshes without relying on such annotations. We show that the performance of our approach is on par with prior work that relies on ground-truth keypoints, and more importantly, we demonstrate the generality of our method by setting new baselines on a larger set of categories from ImageNet - for which keypoints are not available - without any class-specific hyperparameter tuning. We release our code at https://github.com/dariopavllo/textured-3d-gan

****

Multi-Modality Associative Bridging Through Memory: Speech Sound Recollected From Face Video

Minsu Kim, Joanna Hong, Se Jin Park, Yong Man Ro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 296-306

In this paper, we introduce a novel audio-visual multi-modal bridging framework that can utilize both audio and visual information, even with uni-modal inputs. We exploit a memory network that stores source (i.e., visual) and target (i.e., audio) modal representations, where source modal representation is what we are given, and target modal representations are what we want to obtain from the memory network. We then construct an associative bridge between source and target memories that considers the interrelationship between the two memories. By learning the interrelationship through the associative bridge, the proposed bridging fra

mework is able to obtain the target modal representations inside the memory netw
ork, even with the source modal input only, and it provides rich information for
 its downstream tasks. We apply the proposed framework to two tasks: lip reading
 and speech reconstruction from silent video. Through the proposed associative b
ridge and modality-specific memories, each task knowledge is enriched with the r
ecalled audio context, achieving state-of-the-art performance. We also verify th
at the associative bridge properly relates the source and target memories.
****

Warp Consistency for Unsupervised Learning of Dense Correspondences
Prune Truong, Martin Danelljan, Fisher Yu, Luc Van Gool; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10346-10356
The key challenge in learning dense correspondences lies in the lack of ground-t
ruth matches for real image pairs. While photometric consistency losses provide
unsupervised alternatives, they struggle with large appearance changes, which ar
e ubiquitous in geometric and semantic matching tasks. Moreover, methods relying
 on synthetic training pairs often suffer from poor generalisation to real data.
 We propose Warp Consistency, an unsupervised learning objective for dense corre
spondence regression. Our objective is effective even in settings with large app
earance and view-point changes. Given a pair of real images, we first construct
an image triplet by applying a randomly sampled warp to one of the original imag
es. We derive and analyze all flow-consistency constraints arising between the t
riplet. From our observations and empirical results, we design a general unsuper
vised objective employing two of the derived constraints. We validate our warp c
onsistency loss by training three recent dense correspondence networks for the g
eometric and semantic matching tasks. Our approach sets a new state-of-the-art o
n several challenging benchmarks, including MegaDepth, RobotCar and TSS. Code an
d models are at github.com/PruneTruong/DenseMatching.
****

Towards Rotation Invariance in Object Detection
Agastya Kalra, Guy Stoppi, Bradley Brown, Rishav Agarwal, Achuta Kadambi; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 3530-3540
Rotation augmentations generally improve a model's invariance/equivariance to ro
tation - except in object detection. In object detection the shape is not known,
 therefore rotation creates a label ambiguity. We show that the de-facto method
for bounding box label rotation, the Largest Box Method, creates very large labe
ls, leading to poor performance and in many cases worse performance than using n
o rotation at all. We propose a new method of rotation augmentation that can be
implemented in a few lines of code. First, we create a differentiable approximat
ion of label accuracy and show that axis-aligning the bounding box around an ell
ipse is optimal. We then introduce Rotation Uncertainty (RU) Loss, allowing the
model to adapt to the uncertainty of the labels. On five different datasets (inc
luding COCO, PascalVOC, and Transparent Object Bin Picking), this approach impro
ves the rotational invariance of both one-stage and two-stage architectures when
 measured with AP, AP50, and AP75.
****

Unlocking the Potential of Ordinary Classifier: Class-Specific Adversarial Erasi
ng Framework for Weakly Supervised Semantic Segmentation
Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, Kuk-Jin Yoon; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
 pp. 6994-7003
Weakly supervised semantic segmentation (WSSS) using image-level classification
labels usually utilizes the Class Activation Maps (CAMs) to localize objects of
interest in images. While pointing out that CAMs only highlight the most discrim
inative regions of the classes of interest, adversarial erasing (AE) methods hav
e been proposed to further explore the less discriminative regions. In this pape
r, we review the potential of the pre-trained classifier which is trained on the
 raw images. We experimentally verify that the ordinary classifier already has t
he capability to activate the less discriminative regions if the most discrimina
tive regions are erased to some extent. Based on that, we propose a class-specif

ic AE-based framework that fully exploits the potential of an ordinary classifie r. Our framework (1) adopts the ordinary classifier to notify the regions to be erased and (2) generates a class-specific mask for erasing by randomly sampling a single specific class to be erased (target class) among the existing classes o n the image for obtaining more precise CAMs. Specifically, with the guidance of the ordinary classifier, the proposed CAMs Generation Network (CGNet) is enforce d to generate a CAM of the target class while constraining the CAM not to intrud e the object regions of the other classes. Along with the pseudo-labels refined from our CAMs, we achieve the state-of-the-art WSSS performance on both PASCAL V OC 2012 and MS-COCO dataset only with image-level supervision. The code is avail able at https://github.com/KAIST-vilab/OC-CSE.

****

Sparse-to-Dense Feature Matching: Intra and Inter Domain Cross-Modal Learning in Domain Adaptation for 3D Semantic Segmentation

Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, Yulan Guo; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 7108-7117

Domain adaptation is critical for success when confronting with the lack of anno tations in a new domain. As the huge time consumption of labeling process on 3D point cloud, domain adaptation for 3D semantic segmentation is of great expectat ion. With the rise of multi-modal datasets, large amount of 2D images are access ible besides 3D point clouds. In light of this, we propose to further leverage 2 D data for 3D domain adaptation by intra and inter domain cross modal learning. As for intra-domain cross modal learning, most existing works sample the dense 2 D pixel-wise features into the same size with sparse 3D point-wise features, res ulting in the abandon of numerous useful 2D features. To address this problem, w e propose Dynamic sparse-to-dense Cross Modal Learning (DsCML) to increase the s ufficiency of multi-modality information interaction for domain adaptation. For inter-domain cross modal learning, we further advance Cross Modal Adversarial Le arning (CMAL) on 2D and 3D data which contains different semantic content aiming to promote high-level modal complementarity. We evaluate our model under variou s multi-modality domain adaptation settings including day-to-night, country-to-c ountry and dataset-to-dataset, brings large improvements over both uni-modal and multi-modal domain adaptation methods on all settings.

****

Toward Spatially Unbiased Generative Models

Jooyoung Choi, Jungbeom Lee, Yonghyun Jeong, Sungroh Yoon; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14253-14262

Recent image generation models show remarkable generation performance. However, they mirror strong location preference in datasets, which we call spatial bias. Therefore, generators render poor samples at unseen locations and scales. We arg ue that the generators rely on their implicit positional encoding to render spat ial content. From our observations, the generator's implicit positional encoding is translation-variant, making the generator spatially biased. To address this issue, we propose injecting explicit positional encoding at each scale of the ge nerator. By learning the spatially unbiased generator, we facilitate the robust use of generators in multiple tasks, such as GAN inversion, multi-scale generati on, generation of arbitrary sizes and aspect ratios. Furthermore, we show that o ur method can also be applied to denoising diffusion probabilistic models.

****

ReconfigISP: Reconfigurable Camera Image Processing Pipeline

Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, Jinwei Gu; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2021, pp. 4248-4257

Image Signal Processor (ISP) is a crucial component in digital cameras that tran sforms sensor signals into images for us to perceive and understand. Existing IS P designs always adopt a fixed architecture, e.g., several sequential modules co nnected in a rigid order. Such a fixed ISP architecture may be suboptimal for re al-world applications, where camera sensors, scenes and tasks are diverse. In th is study, we propose a novel Reconfigurable ISP (ReconfigISP) whose architecture and parameters can be automatically tailored to specific data and tasks. In par ticular, we implement several ISP modules, and enable backpropagation for each m

odule by training a differentiable proxy, hence allowing us to leverage the popu lar differentiable neural architecture search and effectively search for the opt imal ISP architecture. A proxy tuning mechanism is adopted to maintain the accur acy of proxy networks in all cases. Extensive experiments conducted on image res toration and object detection, with different sensors, light conditions and effi ciency constraints, validate the effectiveness of ReconfigISP. Only hundreds of parameters need tuning for every task.
****

Multi-Expert Adversarial Attack Detection in Person Re-Identification Using Cont ext Inconsistency

Xueping Wang, Shasha Li, Min Liu, Yaonan Wang, Amit K. Roy-Chowdhury; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15097-15107

The success of deep neural networks (DNNs) has promoted the widespread applicati ons of person re-identification (ReID). However, ReID systems inherit the vulner ability of DNNs to malicious attacks of visually inconspicuous adversarial pertu rbations. Detection of adversarial attacks is, therefore, a fundamental requirem ent for robust ReID systems. In this work, we propose a Multi-Expert Adversarial Attack Detection (MEAAD) approach to achieve this goal by checking context inco nsistency, which is suitable for any DNNs-based ReID systems. Specifically, thre e kinds of context inconsistencies caused by adversarial attacks are employed to learn a detector for detecting adversarial attacks, i.e., a) the embedding dist ances between a perturbed query person image and its top-K retrievals are genera lly larger than those between a benign query image and its top-K retrievals, b) the embedding distances among the top-K retrievals of a perturbed query image ar e larger than those of a benign query image, c) the top-K retrievals of a benign query image obtained with multiple expert ReID models tend to be consistent, wh ich is not preserved when attacks are present. Extensive experiments on the Mark et1501 and DukeMTMC-ReID datasets show that, as the first adversarial attack det ection approach for ReID, MEAAD effectively detects various adversarial attacks and achieves high ROC-AUC (over 97.5%).
****

Video Instance Segmentation With a Propose-Reduce Paradigm

Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, Jiaya Jia; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1739-1748

Video instance segmentation (VIS) aims to segment and associate all instances of predefined classes for each frame in videos. Prior methods usually obtain segme ntation for a frame or clip first, and merge the incomplete results by tracking or matching. These methods may cause error accumulation in the merging step. Con trarily, we propose a new paradigm -- Propose-Reduce, to generate complete seque nces for input videos by a single step. We further build a sequence propagation head on the existing image-level instance segmentation network for long-term pro pagation. To ensure robustness and high recall of our proposed framework, multip le sequences are proposed where redundant sequences of the same instance are red uced. We achieve state-of-the-art performance on two representative benchmark da tasets -- we obtain 47.6% in terms of AP on YouTube-VIS validation set and 70.4% for J&F on DAVIS-UVOS validation set.
****

Divide-and-Assemble: Learning Block-Wise Memory for Unsupervised Anomaly Detecti on

Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Hong Zhou; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021 , pp. 8791-8800

Reconstruction-based methods play an important role in unsupervised anomaly dete ction in images. Ideally, we expect a perfect reconstruction for normal samples and poor reconstruction for abnormal samples. Since the generalizability of deep neural networks is difficult to control, existing models such as autoencoder do not work well. In this work, we interpret the reconstruction of an image as a d ivide-and-assemble procedure. Surprisingly, by varying the granularity of divisi on on feature maps, we are able to modulate the reconstruction capability of the

model for both normal and abnormal samples. That is, finer granularity leads to better reconstruction, while coarser granularity leads to poorer reconstruction. With proper granularity, the gap between the reconstruction error of normal and abnormal samples can be maximized. The divide-and-assemble framework is implemented by embedding a novel multi-scale block-wise memory module into an autoencoder network. Besides, we introduce adversarial learning and explore the semantic latent representation of the discriminator, which improves the detection of subtle anomaly. We achieve state-of-the-art performance on the challenging MVTec AD dataset. Remarkably, we improve the vanilla autoencoder model by 10.1% in terms of the AUROC score.

****

Dense Deep Unfolding Network With 3D-CNN Prior for Snapshot Compressive Imaging
Zhuoyuan Wu, Jian Zhang, Chong Mou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4892-4901

Snapshot compressive imaging (SCI) aims to record three-dimensional signals via a two-dimensional camera. For the sake of building a fast and accurate SCI recovery algorithm, we incorporate the interpretability of model-based methods and the speed of learning-based ones and present a novel dense deep unfolding network (DUN) with 3D-CNN prior for SCI, where each phase is unrolled from an iteration of Half-Quadratic Splitting (HQS). To better exploit the spatial-temporal correlation among frames and address the problem of information loss between adjacent phases in existing DUNs, we propose to adopt the 3D-CNN prior in our proximal mapping module and develop a novel dense feature map (DFM) strategy, respectively. Besides, in order to promote network robustness, we further propose a dense feature map adaption (DFMA) module to allow inter-phase information to fuse adaptively. All the parameters are learned in an end-to-end fashion. Extensive experiments on simulation data and real data verify the superiority of our method. The source code is available at \href https://github.com/jianzhangcs/SCI3D https://github.com/jianzhangcs/SCI3D .

****

SelfReg: Self-Supervised Contrastive Regularization for Domain Generalization
Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, Jaekoo Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9619-9628

In general, an experimental environment for deep learning assumes that the training and the test dataset are sampled from the same distribution. However, in real-world situations, a difference in the distribution between two datasets, i.e. domain shift, may occur, which becomes a major factor impeding the generalization performance of the model. The research field to solve this problem is called domain generalization, and it alleviates the domain shift problem by extracting domain-invariant features explicitly or implicitly. In recent studies, contrastive learning-based domain generalization approaches have been proposed and achieved high performance. These approaches require sampling of the negative data pair. However, the performance of contrastive learning fundamentally depends on quality and quantity of negative data pairs. To address this issue, we propose a new regularization method for domain generalization based on contrastive learning, called self-supervised contrastive regularization (SelfReg). The proposed approach uses only positive data pairs, thus it resolves various problems caused by negative pair sampling. Moreover, we propose a class-specific domain perturbation layer (CDPL), which makes it possible to effectively apply mixup augmentation even when only positive data pairs are used. The experimental results show that the techniques incorporated by SelfReg contributed to the performance in a compatible manner. In the recent benchmark, DomainBed, the proposed method shows comparable performance to the conventional state-of-the-art alternatives.

****

Towards the Unseen: Iterative Text Recognition by Distilling From Errors
Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yi-Zhe Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14950-14959

Visual text recognition is undoubtedly one of the most extensively researched to

pics in computer vision. Great progress have been made to date, with the latest models starting to focus on the more practical "in-the-wild" setting. However, a salient problem still hinders practical deployment -- prior arts mostly struggle with recognising unseen (or rarely seen) character sequences. In this paper, we put forward a novel framework to specifically tackle this "unseen" problem. Our framework is iterative in nature, in that it utilises predicted knowledge of character sequences from a previous iteration, to augment the main network in improving the next prediction. Key to our success is a unique cross-modal variational autoencoder to act as a feedback module, which is trained with the presence of textual error distribution data. This module importantly translate a discrete predicted character space, to a continuous affine transformation parameter space used to condition the visual feature map at next iteration. Experiments on common datasets have shown competitive performance over state-of-the-arts under the conventional setting. Most importantly, under the new disjoint setup where train-test labels are mutually exclusive, ours offers the best performance thus showcasing the capability of generalising onto unseen words.

****

SA-ConvONet: Sign-Agnostic Optimization of Convolutional Occupancy Networks

Jiapeng Tang, Jiabao Lei, Dan Xu, Feiying Ma, Kui Jia, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6504-6513

Surface reconstruction from point clouds is a fundamental problem in the computer vision and graphics community. Recent state-of-the-arts solve this problem by individually optimizing each local implicit field during inference. Without considering the geometric relationships between local fields, they typically require accurate normals to avoid the sign conflict problem in overlapped regions of local fields, which severely limits their applicability to raw scans where surface normals could be unavailable. Although SAL breaks this limitation via sign-agnostic learning, further works still need to explore how to extend this technique for local shape modeling. To this end, we propose to learn implicit surface reconstruction by sign-agnostic optimization of convolutional occupancy networks, to simultaneously achieve advanced scalability to large-scale scenes, generality to novel shapes, and applicability to raw scans in a unified framework. Concretely, we achieve this goal by a simple yet effective design, which further optimizes the pre-trained occupancy prediction networks with an unsigned cross-entropy loss during inference. The learning of occupancy fields is conditioned on convolutional features from an hourglass network architecture. Extensive experimental comparisons with previous state-of-the-arts on both object-level and scene-level datasets demonstrate the superior accuracy of our approach for surface reconstruction from un-orientated point clouds. The code is available at https://github.com/tangjiapeng/SA-ConvONet.

****

Rethinking Spatial Dimensions of Vision Transformers

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, Seong Joon Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11936-11945

Vision Transformer (ViT) extends the application range of transformers from language processing to computer vision tasks as being an alternative architecture against the existing convolutional neural networks (CNN). Since the transformer-based architecture has been innovative for computer vision modeling, the design convention towards an effective architecture has been less studied yet. From the successful design principles of CNN, we investigate the role of spatial dimension conversion and its effectiveness on transformer-based architecture. We particularly attend to the dimension reduction principle of CNNs; as the depth increases, a conventional CNN increases channel dimension and decreases spatial dimensions. We empirically show that such a spatial dimension reduction is beneficial to a transformer architecture as well, and propose a novel Pooling-based Vision Transformer (PiT) upon the original ViT model. We show that PiT achieves the improved model capability and generalization performance against ViT. Throughout the extensive experiments, we further show PiT outperforms the baseline on several ta

sks such as image classification, object detection, and robustness evaluation. S
ource codes and ImageNet models are available at https://github.com/naver-ai/pit
.
****

Move2Hear: Active Audio-Visual Source Separation
Sagnik Majumder, Ziad Al-Halah, Kristen Grauman; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 275-285
We introduce the active audio-visual source separation problem, where an agent m
ust move intelligently in order to better isolate the sounds coming from an obje
ct of interest in its environment. The agent hears multiple audio sources simult
aneously (e.g., a person speaking down the hall in a noisy household) and it mus
t use its eyes and ears to automatically separate out the sounds originating fro
m a target object within a limited time budget. Towards this goal, we introduce
a reinforcement learning approach that trains movement policies controlling the
agent's camera and microphone placement over time, guided by the improvement in
predicted audio separation quality. We demonstrate our approach in scenarios mot
ivated by both augmented reality (system is already co-located with the target o
bject) and mobile robotics (agent begins arbitrarily far from the target object)
. Using state-of-the-art realistic audio-visual simulations in 3D environments,
we demonstrate our model's ability to find minimal movement sequences with maxim
al payoff for audio source separation. Project: http://vision.cs.utexas.edu/proj
ects/move2hear
****

VIL-100: A New Dataset and a Baseline Model for Video Instance Lane Detection
Yujun Zhang, Lei Zhu, Wei Feng, Huazhu Fu, Mingqian Wang, Qingxia Li, Cheng Li,
Song Wang; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2021, pp. 15681-15690
Lane detection plays a key role in autonomous driving. While car cameras always
take streaming videos on the way, current lane detection works mainly focus on i
ndividual images (frames) by ignoring dynamics along the video. In this work, we
 collect a new video instance lane detection (VIL-100) dataset, which contains 1
00 videos with in total 10,000 frames, acquired from different real traffic scen
arios. All the frames in each video are manually annotated to a high-quality ins
tance-level lane annotation, and a set of frame-level and video-level metrics ar
e included for quantitative performance evaluation. Moreover, we propose a new b
aseline model, named multi-level memory aggregation network (MMA-Net), for video
 instance lane detection. In our approach, the representation of current frame i
s enhanced by attentively aggregating both local and global memory features from
 other frames. Experiments on the new collected dataset show that the proposed M
MA-Net outperforms state-of-the-art lane detection methods and video object segm
entation methods. We release our dataset and code at https://github.com/yujun0-0
/MMA-Net.
****

Revitalizing Optimization for 3D Human Pose and Shape Estimation: A Sparse Const
rained Formulation
Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, Must
afa Mukadam; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2021, pp. 11457-11466
We propose a novel sparse constrained formulation and from it derive a real-time
 optimization method for 3D human pose and shape estimation. Our optimization me
thod, SCOPE (Sparse Constrained Optimization for 3D human Pose and shapE estimat
ion), is orders of magnitude faster (avg. 4 ms convergence) than existing optimi
zation methods, while being mathematically equivalent to their dense unconstrain
ed formulation under mild assumptions. We achieve this by exploiting the underly
ing sparsity and constraints of our formulation to efficiently compute the Gauss
-Newton direction. We show that this computation scales linearly with the number
 of joints and measurements of a complex 3D human model, in contrast to prior wo
rk where it scales cubically due to their dense unconstrained formulation. Based
 on our optimization method, we present a real-time motion capture framework tha
t estimates 3D human poses and shapes from a single image at over 30 FPS. In ben

chmarks against state-of-the-art methods on multiple public datasets, our framew ork outperforms other optimization methods and achieves competitive accuracy aga inst regression methods. Project page with code and videos: https://sites.google .com/view/scope-human/.

****

AA-RMVSNet: Adaptive Aggregation Recurrent Multi-View Stereo Network

Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, Guoping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6187-6196

In this paper, we present a novel recurrent multi-view stereo network based on l ong short-term memory (LSTM) with adaptive aggregation, namely AA-RMVSNet. We fi rstly introduce an intra-view aggregation module to adaptively extract image fea tures by using context-aware convolution and multi-scale aggregation, which effi ciently improves the performance on challenging regions, such as thin objects an d large low-textured surfaces. To overcome the difficulty of varying occlusion i n complex scenes, we propose an inter-view cost volume aggregation module for ad aptive pixel-wise view aggregation, which is able to preserve better-matched pai rs among all views. The two proposed adaptive aggregation modules are lightweigh t, effective and complementary regarding improving the accuracy and completeness of 3D reconstruction. Instead of conventional 3D CNNs, we utilize a hybrid netw ork with recurrent structure for cost volume regularization, which allows high-r esolution reconstruction and finer hypothetical plane sweep. The proposed networ k is trained end-to-end and achieves excellent performance on various datasets. It ranks 1st among all submissions on Tanks and Temples benchmark and achieves c ompetitive results on DTU dataset, which exhibits strong generalizability and ro bustness. Implementation of our method is available at https://github.com/QT-Zhu /AA-RMVSNet.

****

ME-PCN: Point Completion Conditioned on Mask Emptiness

Bingchen Gong, Yinyu Nie, Yiqun Lin, Xiaoguang Han, Yizhou Yu; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12488-1 2497

Point completion refers to completing the missing geometries of an object from i ncomplete observations. Main-stream methods predict the missing shapes by decodi ng a global feature learned from the input point cloud, which often leads to def icient results in preserving topology consistency and surface details. In this w ork, we present ME-PCN, a point completion network that leverages `emptiness' in 3D shape space. Given a single depth scan, previous methods often encode the oc cupied partial shapes while ignoring the empty regions (e.g. holes) in depth map s. In contrast, we argue that these `emptiness' clues indicate shape boundaries that can be used to improve topology representation and detail granularity on su rfaces. Specifically, our ME-PCN encodes both the occupied point cloud and the n eighboring `empty points'. It estimates coarse-grained but complete and reasonab le surface points in the first stage, followed by a refinement stage to produce fine-grained surface details. Comprehensive experiments verify that our ME-PCN p resents better qualitative and quantitative performance against the state-of-the -art. Besides, we further prove that our `emptiness' design is lightweight and e asy to embed in existing methods, which shows consistent effectiveness in improv ing the CD and EMD scores.

****

Full-Body Motion From a Single Head-Mounted Device: Generating SMPL Poses From P artial Observations

Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J. Cashma n, Jamie Shotton; Proceedings of the IEEE/CVF International Conference on Comput er Vision (ICCV), 2021, pp. 11687-11697

The increased availability and maturity of head-mounted and wearable devices ope ns up opportunities for remote communication and collaboration. However, the sig nal streams provided by these devices (e.g., head pose, hand pose, and gaze dire ction) do not represent a whole person. One of the main open problems is therefo re how to leverage these signals to build faithful representations of the user.

In this paper, we propose a method based on variational autoencoders to generate articulated poses of a human skeleton based on noisy streams of head and hand pose. Our approach relies on a model of pose likelihood that is novel and theoretically well-grounded. We demonstrate on publicly available datasets that our method is effective even from very impoverished signals and investigate how pose prediction can be made more accurate and realistic.

****

LOKI: Long Term and Key Intentions for Trajectory Prediction

Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, Chiho Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9803-9812

Recent advances in trajectory prediction have shown that explicit reasoning about agents' intent is important to accurately forecast their motion. However, the current research activities are not directly applicable to intelligent and safety critical systems. This is mainly because very few public datasets are available, and they only consider pedestrian-specific intents for a short temporal horizon from a restricted egocentric view. To this end, we propose LOKI (LOng term and Key Intentions), a novel large-scale dataset that is designed to tackle joint trajectory and intention prediction for heterogeneous traffic agents (pedestrians and vehicles) in an autonomous driving setting. The LOKI dataset is created to discover several factors that may affect intention, including i) agent's own will, ii) social interactions, iii) environmental constraints, and iv) contextual information. We also propose a model that jointly performs trajectory and intention prediction, showing that recurrently reasoning about intention can assist with trajectory prediction. We show our method outperforms state-of-the-art trajectory prediction methods by upto 27% and also provide a baseline for frame-wise intention estimation. The dataset is available at https://usa.honda-ri.com/loki

****

Three Steps to Multimodal Trajectory Prediction: Modality Clustering, Classification and Synthesis

Jianhua Sun, Yuxuan Li, Hao-Shu Fang, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13250-13259

Multimodal prediction results are essential for trajectory prediction task as there is no single correct answer for the future. Previous frameworks can be divided into three categories: regression, generation and classification frameworks. However, these frameworks have weaknesses in different aspects so that they cannot model the multimodal prediction task comprehensively. In this paper, we present a novel insight along with a brand-new prediction framework by formulating multimodal prediction into three steps: modality clustering, classification and synthesis, and address the shortcomings of earlier frameworks. Exhaustive experiments on popular benchmarks have demonstrated that our proposed method surpasses state-of-the-art works even without introducing social and map information. Specifically, we achieve 19.2% and 20.8% improvement on ADE and FDE respectively on ETH/UCY dataset.

****

Orthogonal Jacobian Regularization for Unsupervised Disentanglement in Image Generation

Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, Wangmeng Zuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6721-6730

Unsupervised disentanglement learning is a crucial issue for understanding and exploiting deep generative models. Recently, SeFa tries to find latent disentangled directions by performing SVD on the first projection of a pre-trained GAN. However, it is only applied to the first layer and works in a post-processing way. Hessian Penalty minimizes the off-diagonal entries of the output's Hessian matrix to facilitate disentanglement, and can be applied to multi-layers. However, it constrains each entry of output independently, making it not sufficient in disentangling the latent directions (e.g., shape, size, rotation, etc.) of spatially correlated variations. In this paper, we propose a simple Orthogonal Jacobian Regularization (OroJaR) to encourage deep generative model to learn disentangled

representations. It simply encourages the variation of output caused by perturbations on different latent dimensions to be orthogonal, and the Jacobian with respect to the input is calculated to represent this variation. We show that our OroJaR also encourages the output's Hessian matrix to be diagonal in an indirect manner. In contrast to the Hessian Penalty, our OroJaR constrains the output in a holistic way, making it very effective in disentangling latent dimensions corresponding to spatially correlated variations. Quantitative and qualitative experimental results show that our method is effective in disentangled and controllable image generation, and performs favorably against the state-of-the-art methods. Our code is available at https://github.com/csyxwei/OroJaR.

****

SGMNet: Learning Rotation-Invariant Point Cloud Representations via Sorted Gram Matrix

Jianyun Xu, Xin Tang, Yushi Zhu, Jie Sun, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10468-10477

Recently, various works that attempted to introduce rotation invariance to point cloud analysis have devised point-pair features, such as angles and distances. In these methods, however, the point-pair is only comprised of the center point and its adjacent points in a vicinity, which may bring information loss to the local feature representation. In this paper, we instead connect each point densely with all other points in a local neighborhood to compose the point-pairs. Specifically, we present a simple but effective local feature representation, called sorted Gram matrix(SGM), which is not only invariant to arbitrary rotations, but also models the pair-wise relationship of all the points in a neighborhood. In more detail, we utilize vector inner product to model distance- and angle-information between two points, and in a local patch it naturally forms a Gram matrix. In order to guarantee permutation invariance, we sort the correlation value in Gram matrix for each point, therefore this geometric feature names sorted Gram matrix. Furthermore, we mathematically prove that the Gram matrix is rotation-invariant and sufficient to model the inherent structure of a point cloud patch. We then use SGM as features in convolution, which can be readily integrated as a drop-in module into any point-based networks. Finally, we evaluated the proposed method on two widely used datasets, and it outperforms previous state-of-the-arts on both shape classification and part segmentation tasks by a large margin.

****

Instance-Wise Hard Negative Example Generation for Contrastive Learning in Unpaired Image-to-Image Translation

Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14020-14029

Contrastive learning shows great potential in unpaired image-to-image translation, but sometimes the translated results are in poor quality and the contents are not preserved consistently. In this paper, we uncover that the negative examples play a critical role in the performance of contrastive learning for image translation. The negative examples in previous methods are randomly sampled from the patches of different positions in the source image, which are not effective to push the positive examples close to the query examples. To address this issue, we present instance-wise hard Negative Example Generation for Contrastive learning in Unpaired image-to-image Translation (NEGCUT). Specifically, we train a generator to produce negative examples online. The generator is novel from two perspectives: 1) it is instance-wise which means that the generated examples are based on the input image, and 2) it can generate hard negative examples since it is trained with an adversarial loss. With the generator, the performance of unpaired image-to-image translation is significantly improved. Experiments on three benchmark datasets demonstrate that the proposed NEGCUT framework achieves state-of-the-art performance compared to previous methods.

****

Multi-Target Adversarial Frameworks for Domain Adaptation in Semantic Segmentation

Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, Patrick Pérez; Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9072-9081
In this work, we address the task of unsupervised domain adaptation (UDA) for semantic segmentation in presence of multiple target domains: the objective is to train a single model that can handle all these domains at test time. Such a multi-target adaptation is crucial for a variety of scenarios that real-world autonomous systems must handle. It is a challenging set-up since one faces not only the domain gap between the labeled source set and the unlabeled target set, but also the distribution shifts existing within the latter among the different target domains. To this end, we introduce two adversarial frameworks: (i) multi-discriminator, which explicitly aligns each target domain to its counterparts, and (ii) multi-target knowledge transfer, which learns a target-agnostic model thanks to a multi-teacher/single-student distillation mechanism. The evaluation is done on four newly proposed multi-target benchmarks for UDA in semantic segmentation. In all tested scenarios, our approaches consistently outperform baselines, setting competitive standards for the novel task.
****

SSH: A Self-Supervised Framework for Image Harmonization
Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4832-4841
Image harmonization aims to improve the quality of image compositing by matching the "appearance"" (e.g., color tone, brightness and contrast) between foreground and background images. However, collecting large-scale annotated datasets for this task requires complex professional retouching. Instead, we propose a novel Self-Supervised Harmonization framework (SSH) that can be trained using just "free"" natural images without being edited. We reformulate the image harmonization problem from a representation fusion perspective, which separately processes the foreground and background examples, to address the background occlusion issue. This framework design allows for a dual data augmentation method, where diverse [foreground, background, pseudo GT] triplets can be generated by cropping an image with perturbations using 3D color lookup tables (LUTs). In addition, we build a real-world harmonization dataset as carefully created by expert users, for evaluation and benchmarking purposes. Our results show that the proposed self-supervised method outperforms previous state-of-the-art methods in terms of reference metrics, visual quality, and subject user study. Code and dataset will be publicly available.
****

Context-Aware Scene Graph Generation With Seq2Seq Transformers
Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W. Taylor, Maksims Volkovs; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15931-15941
Scene graph generation is an important task in computer vision aimed at improving the semantic understand- ing of the visual world. In this task, the model needs to detect objects and predict visual relationships between them. Most of the existing models predict relationships in parallel assuming their independence. While there are differ- ent ways to capture these dependencies, we explore a conditional approach motivated by the sequence-to-sequence (Seq2Seq) formalism. Different from the previous research, our proposed model predicts visual relationships one at a time in an autoregressive manner by explicitly conditioning on the already predicted relationships. Drawing from translation models in NLP, we propose an encoder- decoder model built using Transformers where the encoder captures global context and long range interactions. The decoder then makes sequential predictions by conditioning on the scene graph constructed so far. In addition, we introduce a novel reinforcement learning-based training strategy tailored to Seq2Seq scene graph generation. By using a self-critical policy gradient training approach with Monte Carlo search we directly optimize for the (mean) recall metrics and bridge the gap between training and evaluation. Experimental results on two public benchmark datasets demonstrate that our Seq2Seq learning approach achieves strong empirical performance, out- performing previous state-of-the-art, while remaining efficient in terms of training and inference time. Full code for

this work is available here: https://github.com/ layer6ai-labs/SGG-Seq2Seq.
****

Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective
Yi Zeng, Won Park, Z. Morley Mao, Ruoxi Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16473-16481
Backdoor attacks have been considered a severe security threat to deep learning. Such attacks can make models perform abnormally on inputs with predefined triggers and still retain state-of-the-art performance on clean data. While backdoor attacks have been thoroughly investigated in the image domain from both attackers' and defenders' sides, an analysis in the frequency domain has been missing thus far. This paper first revisits existing backdoor triggers from a frequency perspective and performs a comprehensive analysis. Our results show that many current backdoor attacks exhibit severe high-frequency artifacts, which persist across different datasets and resolutions. We further demonstrate these high-frequency artifacts enable a simple way to detect existing backdoor triggers at a detection rate of 98.50% without prior knowledge of the attack details and the target model. Acknowledging previous attacks' weaknesses, we propose a practical way to create smooth backdoor triggers without high-frequency artifacts and study their detectability. We show that existing defense works can benefit by incorporating these smooth triggers into their design consideration. Moreover, we show that the detector tuned over stronger smooth triggers can generalize well to unseen weak smooth triggers. In short, our work emphasizes the importance of considering frequency analysis when designing both backdoor attacks and defenses in deep learning.
****

Learning Multi-Scene Absolute Pose Regression With Transformers
Yoli Shavit, Ron Ferens, Yosi Keller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2733-2742
Absolute camera pose regression methods estimate the position and orientation of a camera by only using the captured image. A convolutional backbone with a multi-layer perceptron head is trained with images and pose labels to embed a single reference scene at a time. Recently, this framework was extended for learning multiple scenes with a single model by adding a multi-layer perceptron head per scene. In this work, we propose to learn multi-scene absolute camera pose regression with transformers, where encoders are used to aggregate activation maps with self-attention and deocoders transform latent features into candidate pose predictions in parallel, each associated with a different scene. This formulation allows our model to focus on general features that are informative for localization while embedding multiple scenes at once. We evaluate our method on commonly benchmarked indoor and outdoor datasets and show that it surpasses both multi-scene and single-scene absolute pose regressors.
****

Self-Supervised 3D Face Reconstruction via Conditional Estimation
Yandong Wen, Weiyang Liu, Bhiksha Raj, Rita Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13289-13298
We present a conditional estimation (CEST) framework to learn 3D facial parameters from 2D single-view images by self-supervised training from videos. CEST is based on the process of analysis by synthesis, where the 3D facial parameters (shape, reflectance, viewpoint, and illumination) are estimated from the face image, and then recombined to reconstruct the 2D face image. In order to learn semantically meaningful 3D facial parameters without explicit access to their labels, CEST couples the estimation of different 3D facial parameters by taking their statistical dependency into account. Specifically, the estimation of any 3D facial parameter is not only conditioned on the given image, but also on the facial parameters that have already been derived. Moreover, the reflectance symmetry and consistency among the video frames are adopted to improve the disentanglement of facial parameters. Together with a novel strategy for incorporating the reflectance symmetry and consistency, CEST can be efficiently trained with in-the-wild video clips. Both qualitative and quantitative experiments demonstrate the effectiveness of CEST.

****
Training Multi-Object Detector by Estimating Bounding Box Distribution for Input Image
Jaeyoung Yoo, Hojun Lee, Inseop Chung, Geonseok Seo, Nojun Kwak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3437-3446
In multi-object detection using neural networks, the fundamental problem is, "How should the network learn a variable number of bounding boxes in different input images?". Previous methods train a multi-object detection network through a procedure that directly assigns the ground truth bounding boxes to the specific locations of the network's output. However, this procedure makes the training of a multi-object detection network too heuristic and complicated. In this paper, we reformulate the multi-object detection task as a problem of density estimation of bounding boxes. Instead of assigning each ground truth to specific locations of network's output, we train a network by estimating the probability density of bounding boxes in an input image using a mixture model. For this purpose, we propose a novel network for object detection called Mixture Density Object Detector (MDOD), and the corresponding objective function for the density-estimation-based training. We applied MDOD to MS COCO dataset. Our proposed method not only deals with multi-object detection problems in a new approach, but also improves detection performances through MDOD. The code is available: https://github.com/yoojy31/MDOD.
****
Defocus Map Estimation and Deblurring From a Single Dual-Pixel Image
Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T. Barron, Pratul P. Srinivasan, Jiawen Chen, Ioannis Gkioulekas, Rahul Garg; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2228-2238
We present a method that takes as input a single dual-pixel image, and simultaneously estimates the image's defocus map---the amount of defocus blur at each pixel---and recovers an all-in-focus image. Our method is inspired from recent works that leverage the dual-pixel sensors available in many consumer cameras to assist with autofocus, and use them for recovery of defocus maps or all-in-focus images. These prior works have solved the two recovery problems independently of each other, and often require large labeled datasets for supervised training. By contrast, we show that it is beneficial to treat these two closely-connected problems simultaneously. To this end, we set up an optimization problem that, by carefully modeling the optics of dual-pixel images, jointly solves both problems. We use data captured with a consumer smartphone camera to demonstrate that, after a one-time calibration step, our approach improves upon prior works for both defocus map estimation and blur removal, despite being entirely unsupervised.
****
DivAug: Plug-In Automated Data Augmentation With Explicit Diversity Maximization
Zirui Liu, Haifeng Jin, Ting-Hsiang Wang, Kaixiong Zhou, Xia Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4762-4770
Human-designed data augmentation strategies havebeen replaced by automatically learned augmentation pol-icy in the past two years. Specifically, recent works haveexperimentally shown that the superior performance of theautomated methods stems from increasing the diversity ofaugmented data. However, two factors regard-ing the diversity of augmented data are still missing: 1)the explicit definition (and thus measurement) of diversityand 2) the quantifiable relationship between diversity andits regularization effects. To fill this gap, we propose a di-versity measure called "Variance Diversity" and theoreti-cally show that the regularization effect of data augmenta-tion is promised by Variance Diversity. We confirm in exper-iments that the relative gain from automated data augmen-tation in test accuracy of a given model is highly correlatedto Variance Diversity. To improve the search process ofautomated augmentation, an unsupervised sampling-basedframework,DivAug, is designed to directly optimize Vari-ance Diversity and hence strengthen the regularization ef-fect. Without requiring a separate search process, the per-formance gain from DivAug is comparable with state-of-the-art method

with better efficiency. Moreover, under thesemi-supervised setting, our framewor k can further improvethe performance of semi-supervised learning algorithmsbased on RandAugment, making it highly applicable to real-world problems, where label ed data is scarce. The code is available at https://github.com/warai-0toko/DivAu g.

****

VMNet: Voxel-Mesh Network for Geodesic-Aware 3D Semantic Segmentation

Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyua n Sun, Hongbo Fu, Chiew-Lan Tai; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 15488-15498

In recent years, sparse voxel-based methods have become the state-of-the-arts fo r 3D semantic segmentation of indoor scenes, thanks to the powerful 3D CNNs. Nev ertheless, being oblivious to the underlying geometry, voxel-based methods suffe r from ambiguous features on spatially close objects and struggle with handling complex and irregular geometries due to the lack of geodesic information. In vie w of this, we present Voxel-Mesh Network (VMNet), a novel 3D deep architecture t hat operates on the voxel and mesh representations leveraging both the Euclidean and geodesic information. Intuitively, the Euclidean information extracted from voxels can offer contextual cues representing interactions between nearby objec ts, while the geodesic information extracted from meshes can help separate objec ts that are spatially close but have disconnected surfaces. To incorporate such information from the two domains, we design an intra-domain attentive module for effective feature aggregation and an inter-domain attentive module for adaptive feature fusion. Experimental results validate the effectiveness of VMNet: speci fically, on the challenging ScanNet dataset for large-scale segmentation of indo or scenes, it outperforms the state-of-the-art SparseConvNet and MinkowskiNet (7 4.6% vs 72.5% and 73.6% in mIoU) with a simpler network structure (17M vs 30M an d 38M parameters). Code release: https://github.com/hzykent/VMNet

****

FMODetect: Robust Detection of Fast Moving Objects

Denys Rozumnyi, Ji■í Matas, Filip Šroubek, Marc Pollefeys, Martin R. Oswald; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202 1, pp. 3541-3549

We propose the first learning-based approach for fast moving objects detection. Such objects are highly blurred and move over large distances within one video f rame. Fast moving objects are associated with a deblurring and matting problem, also called deblatting. We show that the separation of deblatting into consecuti ve matting and deblurring allows achieving real-time performance, i.e. an order of magnitude speed-up, and thus enabling new classes of application. The propose d method detects fast moving objects as a truncated distance function to the tra jectory by learning from synthetic data. For the sharp appearance estimation and accurate trajectory estimation, we propose a matting and fitting network that e stimates the blurred appearance without background, followed by an energy minimi zation based deblurring. The state-of-the-art methods are outperformed in terms of recall, precision, trajectory estimation, and sharp appearance reconstruction . Compared to other methods, such as deblatting, the inference is of several ord ers of magnitude faster and allows applications such as real-time fast moving ob ject detection and retrieval in large video collections.

****

VideoLT: Large-Scale Long-Tailed Video Recognition

Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, Larr y S. Davis; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 7960-7969

Label distributions in real-world are oftentimes long-tailed and imbalanced, res ulting in biased models towards dominant labels. While long-tailed recognition h as been extensively studied for image classification tasks, limited effort has b een made for video domain. In this paper, we introduce VideoLT, a large-scale lo ng-tailed video recognition dataset, as a step toward real-world video recogniti on. VideoLT contains 256,218 untrimmed videos, annotated into 1,004 classes with a long-tailed distribution. Through extensive studies, we demonstrate that stat

e-of-the-art methods used for long-tailed image recognition do not perform well in the video domain due to the additional temporal dimension in video data. This motivates us to propose FrameStack, a simple yet effective method for long-tailed video recognition task. In particular, FrameStack performs sampling at the frame-level in order to balance class distributions, and the sampling ratio is dynamically determined using knowledge derived from the network during training. Experimental results demonstrate that FrameStack can improve classification performance without sacrificing overall accuracy. Code and dataset are available at: https://github.com/17Skye17/VideoLT.
****

Self-Supervised Monocular Depth Estimation for All Day Images Using Domain Separation
Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, Liangjun Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12737-12746
Remarkable results have been achieved by DCNN based self-supervised depth estimation approaches. However, most of these approaches can only handle either day-time or night-time images, while their performance degrades for all-day images due to large domain shift and the variation of illumination between day and night images. To relieve these limitations, we propose a domain-separated network for self-supervised depth estimation of all-day images. Specifically, to relieve the negative influence of disturbing terms (illumination, etc.), we partition the information of day and night image pairs into two complementary sub-spaces: private and invariant domains, where the former contains the unique information (illumination, etc.) of day and night images and the latter contains essential shared information (texture, etc.). Meanwhile, to guarantee that the day and night images contain the same information, the domain-separated network takes the day-time images and corresponding night-time images (generated by GAN) as input, and the private and invariant feature extractors are learned by orthogonality and similarity loss, where the domain gap can be alleviated, thus better depth maps can be expected. Meanwhile, the reconstruction and photometric losses are utilized to estimate complementary information and depth maps effectively. Experimental results demonstrate that our approach achieves state-of-the-art depth estimation results for all-day images on the challenging Oxford RobotCar dataset, proving the superiority of our proposed approach.
****

An Empirical Study of Training Self-Supervised Vision Transformers
Xinlei Chen, Saining Xie, Kaiming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9640-9649
This paper does not describe a novel method. Instead, it studies a straightforward, incremental, yet must-know baseline given the recent progress in computer vision: self-supervised learning for Vision Transformers (ViT). While the training recipes for standard convolutional networks have been highly mature and robust, the recipes for ViT are yet to be built, especially in the self-supervised scenarios where training becomes more challenging. In this work, we go back to basics and investigate the effects of several fundamental components for training self-supervised ViT. We observe that instability is a major issue that degrades accuracy, and it can be hidden by apparently good results. We reveal that these results are indeed partial failure, and they can be improved when training is made more stable. We benchmark ViT results in MoCo v3 and several other self-supervised frameworks, with ablations in various aspects. We discuss the currently positive evidence as well as challenges and open questions. We hope that this work will provide useful data points and experience for future research.
****

RangeDet: In Defense of Range View for LiDAR-Based 3D Object Detection
Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, ZhaoXiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2918-2927
In this paper, we propose an anchor-free single-stage LiDAR-based 3D object detector -- RangeDet. The most notable difference with previous works is that our me

thod is purely based on the range view representation. Compared with the commonl
y used voxelized or Bird's Eye View (BEV) representations, the range view repres
entation is more compact and without quantization error. Although there are work
s adopting it for semantic segmentation, its performance in object detection is
largely behind voxelized or BEV counterparts. We first analyze the existing rang
e-view-based methods and find two issues overlooked by previous works: 1) the sc
ale variation between nearby and far away objects; 2) the inconsistency between
the 2D range image coordinates used in feature extraction and the 3D Cartesian c
oordinates used in output. Then we deliberately design three components to addre
ss these issues in our RangeDet. We test our RangeDet in the large-scale Waymo O
pen Dataset (WOD). Our best model achieves 72.9/75.9/65.8 3D AP on vehicle/pedes
trian/cyclist. These results outperform other range-view-based methods by a larg
e margin, and are overall comparable with the state-of-the-art multi-view-based
methods. Codes will be released at https://github.com/TuSimple/RangeDet.
****

Data-Free Universal Adversarial Perturbation and Black-Box Attack
Chaoning Zhang, Philipp Benz, Adil Karjauv, In So Kweon; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7868-7877
Universal adversarial perturbation (UAP), i.e. a single perturbation to fool the
 network for most images, is widely recognized as a more practical attack becaus
e the UAP can be generated beforehand and applied directly during the attack sta
ge. One intriguing phenomenon regarding untargeted UAP is that most images are m
isclassified to a dominant label. This phenomenon has been reported in previous
works while lacking a justified explanation, for which our work attempts to prov
ide an alternative explanation. For a more practical universal attack, our inves
tigation of untargeted UAP focuses on alleviating the dependence on the original
 training samples, from removing the need for sample labels to limiting the samp
le size. Towards strictly data-free untargeted UAP, our work proposes to exploit
 artificial Jigsaw images as the training samples, demonstrating competitive per
formance. We further investigate the possibility of exploiting the UAP for a dat
a-free black-box attack which is arguably the most practical yet challenging thr
eat model. We demonstrate that there exists optimization-free repetitive pattern
s which can successfully attack deep models. Code is available at https://bit.ly
/3y0ZTIC.
****

Learning To Hallucinate Examples From Extrinsic and Intrinsic Supervision
Liangke Gui, Adrien Bardes, Ruslan Salakhutdinov, Alexander Hauptmann, Martial H
ebert, Yu-Xiong Wang; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2021, pp. 8701-8711
Learning to hallucinate additional examples has recently been shown as a promisi
ng direction to address few-shot learning tasks. This work investigates two impo
rtant yet overlooked natural supervision signals for guiding the hallucination p
rocess -- (i) extrinsic: classifiers trained on hallucinated examples should be
close to strong classifiers that would be learned from a large amount of real ex
amples; and (ii) intrinsic: clusters of hallucinated and real examples belonging
 to the same class should be pulled together, while simultaneously pushing apart
 clusters of hallucinated and real examples from different classes. We achieve (
i) by introducing an additional mentor model on data-abundant base classes for d
irecting the hallucinator, and achieve (ii) by performing contrastive learning b
etween hallucinated and real examples. As a general, model-agnostic framework, o
ur dual mentor- and self-directed (DMAS) hallucinator significantly improves few
-shot learning performance on widely used benchmarks in various scenarios.
****

Multiresolution Deep Implicit Functions for 3D Shape Representation
Zhang Chen, Yinda Zhang, Kyle Genova, Sean Fanello, Sofien Bouaziz, Christian Hä
ne, Ruofei Du, Cem Keskin, Thomas Funkhouser, Danhang Tang; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13087-1309
6
We introduce Multiresolution Deep Implicit Functions (MDIF), a hierarchical repr
esentation that can recover fine geometry detail, while being able to perform gl

obal operations such as shape completion. Our model represents a complex 3D shap e with a hierarchy of latent grids, which can be decoded into different levels o f detail and also achieve better accuracy. For shape completion, we propose late nt grid dropout to simulate partial data in the latent space and therefore defer the completing functionality to the decoder side.This along with our multires d esign significantly improves the shape completion quality under decoder-only lat ent optimization. To the best of our knowledge, MDIF is the first deep implicit function model that can at the same time (1) represent different levels of detai l and allow progressive decoding; (2) support both encoder-decoder inference and decoder-only latent optimization, and fulfill multiple applications; (3) perfor m detailed decoder-only shape completion. Experiments demonstrate its superior p erformance against prior art in various 3D reconstruction tasks.

****

Single-Shot Hyperspectral-Depth Imaging With Learned Diffractive Optics

Seung-Hwan Baek, Hayato Ikoma, Daniel S. Jeon, Yuqi Li, Wolfgang Heidrich, Gordo n Wetzstein, Min H. Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2651-2660

Imaging depth and spectrum have been extensively studied in isolation from each other for decades. Recently, hyperspectral-depth (HS-D) imaging emerges to captu re both information simultaneously by combining two different imaging systems; o ne for depth, the other for spectrum. While being accurate, this combinational a pproach induces increased form factor, cost, capture time, and alignment/registr ation problems. In this work, departing from the combinational principle, we pro pose a compact single-shot monocular HS-D imaging method. Our method uses a diff ractive optical element (DOE), the point spread function of which changes with r espect to both depth and spectrum. This enables us to reconstruct spectrum and d epth from a single captured image. To this end, we develop a differentiable simu lator and a neural-network-based reconstruction method that are jointly optimize d via automatic differentiation. To facilitate learning the DOE, we present a fi rst HS-D dataset by building a benchtop HS-D imager that acquires high-quality g round truth. We evaluate our method with synthetic and real experiments by build ing an experimental prototype and achieve state-of-the-art HS-D imaging results.

****

Discriminative Region-Based Multi-Label Zero-Shot Learning

Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, Mubar ak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8731-8740

Multi-label zero-shot learning (ZSL) is a more realistic counter-part of standar d single-label ZSL since several objects can co-exist in a natural image. Howeve r, the occurrence of multiple objects complicates the reasoning and requires reg ion-specific processing of visual features to preserve their contextual cues. We note that the best existing multi-label ZSL method takes a shared approach towa rds attending to region features with a common set of attention maps for all the classes. Such shared maps lead to diffused attention, which does not discrimina tively focus on relevant locations when the number of classes are large. Moreove r, mapping spatially-pooled visual features to the class semantics leads to inte r-class feature entanglement, thus hampering the classification. Here, we propos e an alternate approach towards region-based discriminability-preserving multi-l abel zero-shot classification. Our approach maintains the spatial resolution to preserve region-level characteristics and utilizes a bi-level attention module (BiAM) to enrich the features by incorporating both region and scene context info rmation. The enriched region-level features are then mapped to the class semanti cs and only their class predictions are spatially pooled to obtain image-level p redictions, thereby keeping the multi-class features disentangled. Our approach sets a new state of the art on two large-scale multi-label zero-shot benchmarks: NUS-WIDE and Open Images. On NUS-WIDE, our approach achieves an absolute gain o f 6.9% mAP for ZSL, compared to the best published results.

****

FaPN: Feature-Aligned Pyramid Network for Dense Image Prediction

Shihua Huang, Zhichao Lu, Ran Cheng, Cheng He; Proceedings of the IEEE/CVF Inter

national Conference on Computer Vision (ICCV), 2021, pp. 864-873
Recent advancements in deep neural networks have made remarkable leap-forwards in dense image prediction. However, the issue of feature alignment remains as neglected by most existing approaches for simplicity. Direct pixel addition between upsampled and local features leads to feature maps with misaligned contexts that, in turn, translate to mis-classifications in prediction, especially on object boundaries. In this paper, we propose a feature alignment module that learns transformation offsets of pixels to contextually align upsampled higher-level features; and another feature selection module to emphasize the lower-level features with rich spatial details. We then integrate these two modules in a top-down pyramidal architecture and present the Feature-aligned Pyramid Network (FaPN). Extensive experimental evaluations on four dense prediction tasks and four datasets have demonstrated the efficacy of FaPN, yielding an overall improvement of 1.2 - 2.6 points in AP / mIoU over FPN when paired with Faster / Mask R-CNN. In particular, our FaPN achieves the state-of-the-art of 56.7% mIoU on ADE20K when integrated within Mask-Former. The code is available from https://github.com/EMI-Group/FaPN.
****

Personalized Trajectory Prediction via Distribution Discrimination
Guangyi Chen, Junlong Li, Nuoxing Zhou, Liangliang Ren, Jiwen Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15580-15589
Trajectory prediction is confronted with the dilemma to capture the multi-modal nature of future dynamics with both diversity and accuracy. In this paper, we propose a distribution discrimination method (DisDis) to predict personalized motion pattern by distinguishing the potential distributions in a self-supervised manner. The key motivation of DisDis is the observation that the motion pattern of each person is personalized due to his/her habit, character, or goal. Specifically, we learn the latent distribution to represent different motion patterns and optimize it by contrastive discrimination. The contrastive distribution discrimination encourages latent distributions to be discriminative. Our method could be seamlessly integrated with existing multi-modal stochastic predictive models as a plug-and-play module to learn the more discriminative latent distribution. To evaluate the latent distribution, we further propose a new metric, probability cumulative minimum distance (PCMD) curve, which cumulatively calculates the minimum distance on the sorted probabilities. Experimental results on the ETH and UCY datasets show the effectiveness of our method.
****

GridToPix: Training Embodied Agents With Minimal Supervision
Unnat Jain, Iou-Jen Liu, Svetlana Lazebnik, Aniruddha Kembhavi, Luca Weihs, Alexander G. Schwing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15141-15151
While deep reinforcement learning (RL) promises freedom from hand-labeled data, great successes, especially for Embodied AI, require significant work to create supervision via carefully shaped rewards. Indeed, without shaped rewards, i.e., with only terminal rewards, present-day Embodied AI results degrade significantly across Embodied AI problems from single-agent Habitat-based PointGoal Navigation (SPL drops from 55 to 0) and two-agent AI2-THOR-based Furniture Moving (success drops from 58% to 1%) to three-agent Google Football-based 3 vs. 1 with Keeper (game score drops from 0.6 to 0.1). As training from shaped rewards doesn't scale to more realistic tasks, the community needs to improve the success of training with terminal rewards. For this we propose GridToPix: 1) train agents with terminal rewards in gridworlds that generically mirror Embodied AI environments, i.e., they are independent of the task; 2) distill the learned policy into agents that reside in complex visual worlds. Despite learning from only terminal rewards with identical models and RL algorithms, GridToPix significantly improves results across tasks: from PointGoal Navigation (SPL improves from 0 to 64) and Furniture Moving (success improves from 1% to 25%) to football gameplay (game score improves from 0.1 to 0.6). GridToPix even helps to improve the results of shaped reward training.

****
On the Robustness of Vision Transformers to Adversarial Examples
Kaleel Mahmood, Rigel Mahmood, Marten van Dijk; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7838-7847
Recent advances in attention-based networks have shown that Vision Transformers can achieve state-of-the-art or near state-of-the-art results on many image classification tasks. This puts transformers in the unique position of being a promising alternative to traditional convolutional neural networks (CNNs). While CNNs have been carefully studied with respect to adversarial attacks, the same cannot be said of Vision Transformers. In this paper, we study the robustness of Vision Transformers to adversarial examples. Our analyses of transformer security is divided into three parts. First, we test the transformer under standard white-box and black-box attacks. Second, we study the transferability of adversarial examples between CNNs and transformers. We show that adversarial examples do not readily transfer between CNNs and transformers. Based on this finding, we analyze the security of a simple ensemble defense of CNNs and transformers. By creating a new attack, the self-attention blended gradient attack, we show that such an ensemble is not secure under a white-box adversary. However, under a black-box adversary, we show that an ensemble can achieve unprecedented robustness without sacrificing clean accuracy. Our analysis for this work is done using six types of white-box attacks and two types of black-box attacks. Our study encompasses multiple Vision Transformers, Big Transfer Models and CNN architectures trained on CIFAR-10, CIFAR-100 and ImageNet.
****
HiFT: Hierarchical Feature Transformer for Aerial Tracking
Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, Yiming Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15457-15466
Most existing Siamese-based tracking methods execute the classification and regression of the target object based on the similarity maps. However, they either employ a single map from the last convolutional layer which degrades the localization accuracy in complex scenarios or separately use multiple maps for decision making, introducing intractable computations for aerial mobile platforms. Thus, in this work, we propose an efficient and effective hierarchical feature transformer (HiFT) for aerial tracking. Hierarchical similarity maps generated by multi-level convolutional layers are fed into the feature transformer to achieve the interactive fusion of spatial (shallow layers) and semantics cues (deep layers). Consequently, not only the global contextual information can be raised, facilitating the target search, but also our end-to-end architecture with the transformer can efficiently learn the interdependencies among multi-level features, thereby discovering a tracking-tailored feature space with strong discriminability. Comprehensive evaluations on four aerial benchmarks have proven the effectiveness of HiFT. Real-world tests on the aerial platform have strongly validated its practicability with a real-time speed. Our code is available at https://github.com/vision4robotics/HiFT.
****
Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet
Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E. H. Tay, Jiashi Feng, Shuicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 558-567
Transformers, which are popular for language modeling, have been explored for solving vision tasks recently, e.g., the Vision Transformer (ViT) for image classification. The ViT model splits each image into a sequence of tokens with fixed length and then applies multiple Transformer layers to model their global relation for classification. However, ViT achieves inferior performance to CNNs when trained from scratch on a midsize dataset like ImageNet. We find it is because: 1) the simple tokenization of input images fails to model the important local structure such as edges and lines among neighboring pixels, leading to low training sample efficiency; 2) the redundant attention backbone design of ViT leads to limited feature richness for fixed computation budgets and limited training samples. To overcome such limitations, we propose a new Tokens-To-Token Vision Transfo

rmer (T2T-ViT), which incorporates 1) a layer-wise Tokens-to-Token (T2T) transfo
rmation to progressively structurize the image to tokens by recursively aggregat
ing neighboring Tokens into one Token (Tokens-to-Token), such that local structu
re represented by surrounding tokens can be modeled and tokens length can be red
uced; 2) an efficient backbone with a deep-narrow structure for vision transform
er motivated by CNN architecture design after empirical study. Notably, T2T-ViT
reduces the parameter count and MACs of vanilla ViT by half, while achieving mor
e than 3.0% improvement when trained from scratch on ImageNet. It also outperfor
ms ResNets and achieves comparable performance with MobileNets by directly train
ing on ImageNet. For example, T2T-ViT with comparable size to ResNet50 (21.5M pa
rameters) can achieve 83.3% top1 accuracy in image resolution 384x384 on ImageNe
t.
****

Teacher-Student Adversarial Depth Hallucination To Improve Face Recognition
Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, Ali Etemad; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
3671-3680
We present the Teacher-Student Generative Adversarial Network (TS-GAN) to genera
te depth images from single RGB images in order to boost the performance of face
 recognition systems. For our method to generalize well across unseen datasets,
we design two components in the architecture, a teacher and a student. The teach
er, which itself consists of a generator and a discriminator, learns a latent ma
pping between input RGB and paired depth images in a supervised fashion. The stu
dent, which consists of two generators (one shared with the teacher) and a discr
iminator, learns from new RGB data with no available paired depth information, f
or improved generalization. The fully trained shared generator can then be used
in runtime to hallucinate depth from RGB for downstream applications such as fac
e recognition. We perform rigorous experiments to show the superiority of TS-GAN
 over other methods in generating synthetic depth images. Moreover, face recogni
tion experiments demonstrate that our hallucinated depth along with the input RG
B images boost performance across various architectures when compared to a singl
e RGB modality by average values of +1.2%, +2.6%, and +2.6% for IIIT-D, EURECOM,
 and LFW datasets respectively. We make our implementation public at: https://gi
thub.com/hardik-uppal/teacher-student-gan.git.
****

Interaction via Bi-Directional Graph of Semantic Region Affinity for Scene Parsi
ng
Henghui Ding, Hui Zhang, Jun Liu, Jiaxin Li, Zijian Feng, Xudong Jiang; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp
. 15848-15858
In this work, we devote to address the challenging problem of scene parsing. Pre
vious methods, though capture context to exploit global clues, handle scene pars
ing as a pixel-independent task. However, it is well known that pixels in an ima
ge are highly correlated with each other, especially those from the same semanti
c region, while treating pixels independently fails to take advantage of such co
rrelations. In this work, we treat each respective region in an image as a whole
, and capture the structure topology as well as the affinity among different reg
ions. To this end, we first divide the entire feature maps to different regions
and extract respective global features from them. Next, we construct a directed
graph whose nodes are regional features, and the edge connecting every two nodes
 is the affinity between the regional features they represent. After that, we tr
ansfer the affinity-aware nodes in the directed graph back to corresponding regi
ons of the image, which helps to model the region dependencies and mitigate unre
alistic results. In addition, to further boost the correlation among pixels, we
propose a region-level loss that evaluates all pixels in a region as a whole and
 motivates the network to learn the exclusive regional feature per class. With t
he proposed approach, we achieves new state-of-the-art segmentation results on P
ASCAL-Context, ADE20K, and COCO-Stuff consistently.
****

Online Multi-Granularity Distillation for GAN Compression

Yuxi Ren, Jie Wu, Xuefeng Xiao, Jianchao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6793-6803
Generative Adversarial Networks (GANs) have witnessed prevailing success in yielding outstanding images, however, they are burdensome to deploy on resource-constrained devices due to ponderous computational costs and hulking memory usage. Although recent efforts on compressing GANs have acquired remarkable results, they still exist potential model redundancies and can be further compressed. To solve this issue, we propose a novel online multi-granularity distillation (OMGD) scheme to obtain lightweight GANs, which contributes to generating high-fidelity images with low computational demands. We offer the first attempt to popularize single-stage online distillation for GAN-oriented compression, where the progressively promoted teacher generator helps to refine the discriminator-free based student generator. Complementary teacher generators and network layers provide comprehensive and multi-granularity concepts to enhance visual fidelity from diverse dimensions. Experimental results on four benchmark datasets demonstrate that OMGD successes to compress 40xMACs and 82.5xparameters on Pix2Pix and CycleGAN, without loss of image quality. It reveals that OMGD provides a feasible solution for the deployment of real-time image translation on resource-constrained devices. Our code and models are made public at: https://github.com/bytedance/OMGD
****

Influence-Balanced Loss for Imbalanced Visual Classification
Seulki Park, Jongin Lim, Younghan Jeon, Jin Young Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 735-744
In this paper, we propose a balancing training method to address problems in imbalanced data learning. To this end, we derive a new loss used in the balancing training phase that alleviates the influence of samples that cause an overfitted decision boundary. The proposed loss efficiently improves the performance of any type of imbalance learning methods. In experiments on multiple benchmark data sets, we demonstrate the validity of our method and reveal that the proposed loss outperforms the state-of-the-art cost-sensitive loss methods. Furthermore, since our loss is not restricted to a specific task, model, or training method, it can be easily used in combination with other recent re-sampling, meta-learning, and cost-sensitive learning methods for class-imbalance problems. Our code is made available.
****

Consistency-Aware Graph Network for Human Interaction Understanding
Zhenhua Wang, Jiajun Meng, Dongyan Guo, Jianhua Zhang, Javen Qinfeng Shi, Shengyong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13369-13378
Compared with the progress made on human activity classification, much less success has been achieved on human interaction understanding (HIU). Apart from the latter task is much more challenging, the main cause is that recent approaches learn human interactive relations via shallow graphical models, which is inadequate to model complicated human interactions. In this paper, we propose a consistency-aware graph network, which combines the representative ability of graph network and the consistency-aware reasoning to facilitate the HIU task. Our network consists of three components, a backbone CNN to extract image features, a factor graph network to learn third-order interactive relations among participants, and a consistency-aware reasoning module to enforce labeling and grouping consistencies. Our key observation is that the consistency-aware-reasoning bias for HIU can be embedded into an energy function, minimizing which delivers consistent predictions. An efficient mean-field inference algorithm is proposed, such that all modules of our network could be trained jointly in an end-to-end manner. Experimental results show that our approach achieves leading performance on three benchmarks. Code will be publicly available.
****

Where2Act: From Pixels to Actions for Articulated 3D Objects
Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6813-6823

One of the fundamental goals of visual perception is to allow agents to meaningfully interact with their environment. In this paper, we take a step towards that long-term goal -- we extract highly localized actionable information related to elementary actions such as pushing or pulling for articulated objects with movable parts. For example, given a drawer, our network predicts that applying a pulling force on the handle opens the drawer. We propose, discuss, and evaluate novel network architectures that given image and depth data, predict the set of actions possible at each pixel, and the regions over articulated parts that are likely to move under the force. We propose a learning-from-interaction framework with an online data sampling strategy that allows us to train the network in simulation (SAPIEN) and generalizes across categories. Check the website for code and data release.

****

Towers of Babel: Combining Images, Language, and 3D Geometry for Learning Multimodal Vision

Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, Noah Snavely; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 428-437

The abundance and richness of Internet photos of landmarks and cities has led to significant progress in 3D vision over the past two decades, including automated 3D reconstructions of the world's landmarks from tourist photos. However, a major source of information available for these 3D-augmented collections---language, e.g., from image captions---has been virtually untapped. In this work, we present WikiScenes, a new, large-scale dataset of landmark photo collections that contains descriptive text in the form of captions and hierarchical category names. WikiScenes forms a new testbed for multimodal reasoning involving images, text, and 3D geometry. We demonstrate the utility of WikiScenes for learning semantic concepts over images and 3D models. Our weakly-supervised framework connects images, 3D structure and semantics---utilizing the strong constraints provided by 3D geometry---to associate semantic concepts to image pixels and points in 3D space.

****

End-to-End Unsupervised Document Image Blind Denoising

Mehrdad J. Gangeh, Marcin Plata, Hamid R. Motahari Nezhad, Nigel P Duffy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7888-7897

Removing noise from scanned pages is a vital step before their submission to optical character recognition (OCR) system. Most available image denoising methods are supervised where the pairs of noisy/clean pages are required. However, this assumption is rarely met in real settings. Besides, there is no single model that can remove various noise types from documents. Here, we propose a unified end-to-end unsupervised deep learning model, for the first time, that can effectively remove multiple types of noise, including salt & pepper noise, blurred and/or faded text, as well as watermarks from documents at various levels of intensity. We demonstrate that the proposed model significantly improves the quality of scanned images and the OCR of the pages on several test datasets.

****

Differentiable Dynamic Wirings for Neural Networks

Kun Yuan, Quanquan Li, Shaopeng Guo, Dapeng Chen, Aojun Zhou, Fengwei Yu, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 327-336

A standard practice of deploying deep neural networks is to apply the same architecture to all the input instances. However, a fixed architecture may not be suitable for different data with high diversity. To boost the model capacity, existing methods usually employ larger convolutional kernels or deeper network layers, which incurs prohibitive computational costs. In this paper, we address this issue by proposing Differentiable Dynamic Wirings (DDW), which learns the instance-aware connectivity that creates different wiring patterns for different instances. 1) Specifically, the network is initialized as a complete directed acyclic graph, where the nodes represent convolutional blocks and the edges represent the connection paths. 2) We generate edge weights by a learnable module, Router, a

nd select the edges whose weights are larger than a threshold, to adjust the con nectivity of the neural network structure. 3) Instead of using the same path of the network, DDW aggregates features dynamically in each node, which allows the network to have more representation power. To facilitate effective training, we further represent the network connectivity of each sample as an adjacency matrix . The matrix is updated to aggregate features in the forward pass, cached in the memory, and used for gradient computing in the backward pass. We validate the e ffectiveness of our approach with several mainstream architectures, including Mo bileNetV2, ResNet, ResNeXt, and RegNet. Extensive experiments are performed on I mageNet classification and COCO object detection, which demonstrates the effecti veness and generalization ability of our approach.
****

A Simple Framework for 3D Lensless Imaging With Programmable Masks
Yucheng Zheng, Yi Hua, Aswin C. Sankaranarayanan, M. Salman Asif; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2603 -2612
Lensless cameras provide a framework to build thin imaging systems by replacing the lens in a conventional camera with an amplitude or phase mask near the senso r. Existing methods for lensless imaging can recover the depth and intensity of the scene, but they require solving computationally-expensive inverse problems. Furthermore, existing methods struggle to recover dense scenes with large depth variations. In this paper, we propose a lensless imaging system that captures a small number of measurements using different patterns on a programmable mask. In this context, we make three contributions. First, we present a fast recovery al gorithm to recover textures on a fixed number of depth planes in the scene. Seco nd, we consider the mask design problem, for programmable lensless cameras, and provide a design template for optimizing the mask patterns with the goal of impr oving depth estimation. Third, we use a refinement network as a post-processing step to identify and remove artifacts in the reconstruction. These modifications are evaluated extensively with experimental results on a lensless camera protot ype to showcase the performance benefits of the optimized masks and recovery alg orithms over the state of the art.
****

Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing
Aiyu Cui, Daniel McKee, Svetlana Lazebnik; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2021, pp. 14638-14647
We proposes a flexible person generation framework called Dressing in Order (DiO r), which supports 2D pose transfer, virtual try-on, and several fashion editing tasks. The key to DiOr is a novel recurrent generation pipeline to sequentially put garments on a person, so that trying on the same garments in different orde rs will result in different looks. Our system can produce dressing effects not a chievable by existing work, including different interactions of garments (e.g., wearing a top tucked into the bottom or over it), as well as layering of multipl e garments of the same type (e.g., jacket over shirt over t-shirt). DiOr explici tly encodes the shape and texture of each garment, enabling these elements to be edited separately. Joint training on pose transfer and inpainting helps with de tail preservation and coherence of generated garments. Extensive evaluations sho w that DiOr outperforms other recent methods like ADGAN in terms of output quali ty, and handles a wide range of editing functions for which there is no direct s upervision.
****

Attack As the Best Defense: Nullifying Image-to-Image Translation GANs via Limit -Aware Adversarial Attack
Chin-Yuan Yeh, Hsi-Wen Chen, Hong-Han Shuai, De-Nian Yang, Ming-Syan Chen; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16188-16197
Due to the great success of image-to-image (Img2Img) translation GANs, many appl ications with ethics issues arise, e.g., DeepFake and DeepNude, presenting a cha llenging problem to prevent the misuse of these techniques. In this work, we tac

kle the problem by a new adversarial attack scheme, namely the Nullifying Attack , which cancels the image translation process and proposes a corresponding frame work, the Limit-Aware Self-Guiding Gradient Sliding Attack (LaS-GSA) under a bla ck-box setting. In other words, by processing the image with the proposed LaS-GS A before publishing, any image translation functions can be nullified, which pre vents the images from malicious manipulations. First, we introduce the limit-awa re RGF and gradient sliding mechanism to estimate the gradient that adheres to t he adversarial limit, i.e., the pixel value limitations of the adversarial examp le. We theoretically prove that our model is able to avoid the error caused by t he projection operation in both the direction and the length. Then, an effective self-guiding prior is extracted solely from the threat model and the target ima ge to efficiently leverage the prior information and guide the gradient estimati on process. Extensive experiments demonstrate that LaS-GSA requires fewer querie s to nullify the image translation process with higher success rates than 4 stat e-of-the-art methods.

****

Benchmark Platform for Ultra-Fine-Grained Visual Categorization Beyond Human Per formance

Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, Shengwu Xiong; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 102 85-10295

Deep learning methods have achieved remarkable success in fine-grained visual ca tegorization. Such successful categorization at sub-ordinate level, e.g., differ ent animal or plant species, however relies heavily on the visual differences th at human can observe and the ground-truths are labelled on the basis of such hum an visual observation. In contrast, few research has been done for visual catego rization at the ultra-fine-grained level, i.e., a granularity where even human e xperts can hardly identify the visual differences or are not yet able to give af firmative labels by inferring observed pattern differences. This paper reports o ur efforts towards mitigating this research gap. We introduce the ultra-fine-gra ined (UFG) image dataset, a large collection of 47,114 images from 3,526 categor ies. All the images in the proposed UFG image dataset are grouped into categorie s with different confirmed cultivar names. In addition, we perform an extensive evaluation of state-of-the-art fine-grained classification methods on the propos ed UFG image dataset as comparative baselines. The proposed UFG image dataset an d evaluation protocols is intended to serve as a benchmark platform that can adv ance research of visual classification from approaching human performance to bey ond human ability, via facilitating benchmark data of artificial intelligence (A I) not to be limited by the labels of human intelligence (HI). The dataset is av ailable online at https://github.com/XiaohanYu-GU/Ultra-FGVC.

****

JEM++: Improved Techniques for Training JEM

Xiulong Yang, Shihao Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6494-6503

Joint Energy-based Model (JEM) is a recently proposed hybrid model that retains strong discriminative power of modern CNN classifiers, while generating samples rivaling the quality of GAN-based approaches. In this paper, we propose a variet y of new training procedures and architecture features to improve JEM's accuracy , training stability, and speed altogether. 1) We propose a proximal SGLD to gen erate samples in the proximity of samples from previous step, which improves the stability. 2) We further treat the approximate maximum likelihood learning of E BM as a multi-step differential game, and extend the YOPO framework to cut out r edundant calculations during backpropagation, which accelerates the training sub stantially. 3) Rather than initializing SGLD chain from random noise, we introdu ce a new informative initialization that samples from a distribution estimated f rom training data. 4) This informative initialization allows us to enable batch normalization in JEM, which further releases the power of modern CNN architectur es for hybrid modeling.

****

Contrast and Classify: Training Robust VQA Models

Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, Harsh Agrawal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1604-1613

Recent Visual Question Answering (VQA) models have shown impressive performance on the VQA benchmark but remain sensitive to small linguistic variations in input questions. Existing approaches address this by augmenting the dataset with question paraphrases from visual question generation models or adversarial perturbations. These approaches use the combined data to learn an answer classifier by minimizing the standard cross-entropy loss. To more effectively leverage augmented data, we build on the recent success in contrastive learning. We propose a novel training paradigm (ConClaT) that optimizes both cross-entropy and contrastive losses. The contrastive loss encourages representations to be robust to linguistic variations in questions while the cross-entropy loss preserves the discriminative power of representations for answer prediction. We find that optimizing both losses -- either alternately or jointly -- is key to effective training. On the VQA-Rephrasings benchmark, which measures the VQA model's answer consistency across human paraphrases of a question, ConClaT improves Consensus Score by 1.63% over an improved baseline. In addition, on the standard VQA 2.0 benchmark, we improve the VQA accuracy by 0.78% overall. We also show that ConClaT is agnostic to the type of data-augmentation strategy used.

****

Photon-Starved Scene Inference Using Single Photon Cameras

Bhavya Goyal, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2512-2521

Scene understanding under low-light conditions is a challenging problem. This is due to the small number of photons captured by the camera and the resulting low signal-to-noise ratio (SNR). Single-photon cameras (SPCs) are an emerging sensing modality that are capable of capturing images with high sensitivity. Despite having minimal read-noise, images captured by SPCs in photon-starved conditions still suffer from strong shot noise, preventing reliable scene inference. We propose photon scale-space -- a collection of high-SNR images spanning a wide range of photons-per-pixel (PPP) levels (but same scene content) as guides to train inference model on low photon flux images. We develop training techniques that push images with different illumination levels closer to each other in feature representation space. The key idea is that having a spectrum of different brightness levels during training enables effective guidance, and increases robustness to shot noise even in extreme noise cases. Based on the proposed approach, we demonstrate, via simulations and real experiments with a SPAD camera, high-performance on various inference tasks such as image classification and monocular depth estimation under ultra low-light, down to <1 PPP.

****

Towards Learning Spatially Discriminative Feature Representations

Chaofei Wang, Jiayu Xiao, Yizeng Han, Qisen Yang, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1326-1335

The backbone of traditional CNN classifier is generally considered as a feature extractor, followed by a linear layer which performs the classification. We propose a novel loss function, termed as CAM-loss, to constrain the embedded feature maps with the class activation maps (CAMs) which indicate the spatially discriminative regions of an image for particular categories. CAM-loss drives the backbone to express the features of target category and suppress the features of non-target categories or background, so as to obtain more discriminative feature representations. It can be simply applied in any CNN architecture with neglectable additional parameters and calculations. Experimental results show that CAM-loss is applicable to a variety of network structures and can be combined with mainstream regularization methods to improve the performance of image classification. The strong generalization ability of CAM-loss is validated in the transfer learning and few shot learning tasks. Based on CAM-loss, we also propose a novel CAAM-CAM matching knowledge distillation method. This method directly uses the CAM generated by the teacher network to supervise the CAAM generated by the student n

etwork, which effectively improves the accuracy and convergence rate of the student network.
****

Pyramid Spatial-Temporal Aggregation for Video-Based Person Re-Identification
Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, Dong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12026-12035
Video-based person re-identification aims to associate the video clips of the same person across multiple non-overlapping cameras. Spatial-temporal representations can provide richer and complementary information between frames, which are crucial to distinguish the target person when occlusion occurs. This paper proposes a novel Pyramid Spatial-Temporal Aggregation (PSTA) framework to aggregate the frame-level features progressively and fuse the hierarchical temporal features into a final video-level representation. Thus, short-term and long-term temporal information could be well exploited by different hierarchies. Furthermore, a Spatial-Temporal Aggregation Module (STAM) is proposed to enhance the aggregation capability of PSTA. It mainly consists of two novel attention blocks: Spatial Reference Attention (SRA) and Temporal Reference Attention (TRA). SRA explores the spatial correlations within a frame to determine the attention weight of each location. While TRA extends SRA with the correlations between adjacent frames, temporal consistency information can be fully explored to suppress the interference features and strengthen the discriminative ones. Extensive experiments on several challenging benchmarks demonstrate the effectiveness of the proposed PSTA, and our full model reaches 91.5% and 98.3% Rank-1 accuracy on MARS and DukeMTMC-VID benchmarks.
****

Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation
Yukun Su, Ruizhou Sun, Guosheng Lin, Qingyao Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7004-7014
Data augmentation is vital for deep learning neural networks. By providing massive training samples, it helps to improve the generalization ability of the model. Weakly supervised semantic segmentation (WSSS) is a challenging problem that has been deeply studied in recent years, conventional data augmentation approaches for WSSS usually employ geometrical transformations, random cropping, and color jittering. However, merely increasing the same contextual semantic data does not bring much gain to the networks to distinguish the objects, e.g., the correct image-level classification of "aeroplane" may be not only due to the recognition of the object itself but also its co-occurrence context like "sky", which will cause the model to focus less on the object features. To this end, we present a Context Decoupling Augmentation (CDA) method, to change the inherent context in which the objects appear and thus drive the network to remove the dependence between object instances and contextual information. To validate the effectiveness of the proposed method, extensive experiments on PASCAL VOC 2012 dataset with several alternative network architectures demonstrate that CDA can boost various popular WSSS methods to the new state-of-the-art by a large margin.
****

CAPTRA: CAtegory-Level Pose Tracking for Rigid and Articulated Objects From Point Clouds
Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13209-13218
In this work, we tackle the problem of category-level online pose tracking for objects from point cloud sequences. For the first time, we propose a unified framework that can handle 9DoF object pose tracking for novel rigid object instances as well as per-part pose tracking for articulated objects from known categories. Here the 9DoF pose, comprising 6D pose and 3D size, is equivalent to a 3D amodal bounding box representation with free 6D pose. Given the depth point cloud at the current frame and the estimated pose from the last frame, our novel end-to-end pipeline learns to accurately update the pose. Our pipeline is composed of three modules: 1) a pose canonicalization module that normalizes the pose of the

input depth point cloud; 2) RotationNet, a module that directly regresses small interframe delta rotations; and 3) CoordinateNet, a module that predicts the normalized coordinates and segmentation, enabling analytical computation of the 3D size and translation. Leveraging the small pose regime in the pose-canonicalized point clouds, our method integrates the best of both worlds by combining dense coordinate prediction and direct rotation regression, thus yielding an end-to-end differentiable pipeline optimized for 9DoF pose accuracy (without using non-differentiable RANSAC). Our extensive experiments demonstrate that our method achieves new state-of-the-art performance on category-level rigid object pose and articulated object pose benchmarks at the fastest FPS 12.
****

X-World: Accessibility, Vision, and Autonomy Meet
Jimuyang Zhang, Minglan Zheng, Matthew Boyd, Eshed Ohn-Bar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9762-9771
An important issue facing vision-based intelligent systems today is the lack of accessibility-aware development. A main reason for this issue is the absence of any large-scale, standardized vision benchmarks that incorporate relevant tasks and scenarios related to people with disabilities. This lack of representation hinders even preliminary analysis with respect to underlying pose, appearance, and occlusion characteristics of diverse pedestrians. What is the impact of significant occlusion from a wheelchair on instance segmentation quality? How can interaction with mobility aids, e.g., a long and narrow walking cane, be recognized robustly? To begin addressing such questions, we introduce X-World, an accessibility-centered development environment for vision-based autonomous systems. We tackle inherent data scarcity by leveraging a simulation environment to spawn dynamic agents with various mobility aids. The simulation supports generation of ample amounts of finely annotated, multi-modal data in a safe, cheap, and privacy-preserving manner. Our analysis highlights novel challenges introduced by our benchmark and tasks, as well as numerous opportunities for future developments. We further broaden our analysis using a complementary real-world evaluation benchmark of in-situ navigation by pedestrians with disabilities. Our contributions provide an initial step towards widespread deployment of vision-based agents that can perceive and model the interaction needs of diverse people with disabilities.
****

Target Adaptive Context Aggregation for Video Scene Graph Generation
Yao Teng, Limin Wang, Zhifeng Li, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13688-13697
This paper deals with a challenging task of video scene graph generation (VidSGG), which could serve as a structured video representation for high-level understanding tasks. We present a new detect-to-track paradigm for this task by decoupling the context modeling for relation prediction from the complicated low-level entity tracking. Specifically, we design an efficient method for frame-level VidSGG, termed as Target Adaptive Context Aggregation Network (TRACE), with a focus on capturing spatio-temporal context information for relation recognition. Our TRACE framework streamlines the VidSGG pipeline with a modular design, and presents two unique blocks of Hierarchical Relation Tree (HRTree) construction and Target-adaptive Context Aggregation. More specific, our HRTree first provides an adpative structure for organizing possible relation candidates efficiently, and guides context aggregation module to effectively capture spatio-temporal structure information. Then, we obtain a contextualized feature representation for each relation candidate and build a classification head to recognize its relation category. Finally, we provide a simple temporal association strategy to track TRACE detected results to yield the video-level VidSGG. We perform experiments on two VidSGG benchmarks: ImageNet-VidVRD and Action Genome, and the results demonstrate that our TRACE achieves the state-of-the-art performance. The code and models are made available at https://github.com/MCG-NJU/TRACE.
****

Learnable Boundary Guided Adversarial Training
Jiequan Cui, Shu Liu, Liwei Wang, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15721-15730

Previous adversarial training raises model robustness under the compromise of ac
curacy on natural data. In this paper, we reduce natural accuracy degradation. W
e use the model logits from one clean model to guide learning of another one rob
ust model, taking into consideration that logits from the well trained clean mod
el embed the most discriminative features of natural data, e.g., generalizable c
lassifier boundary. Our solution is to constrain logits from the robust model th
at takes adversarial examples as input and makes it similar to those from the cl
ean model fed with corresponding natural data. It lets the robust model inherit
the classifier boundary of the clean model. Moreover, we observe such boundary g
uidance can not only preserve high natural accuracy but also benefit model robus
tness, which gives new insights and facilitates progress for the adversarial com
munity. Finally, extensive experiments on CIFAR-10, CIFAR-100, and Tiny ImageNet
 testify to the effectiveness of our method. We achieve new state-of-the-art rob
ustness on CIFAR-100 without additional real or synthetic data with auto-attack
benchmark. Our code is available at https://github.com/dvlab-research/LBGAT.
****

Memory-Augmented Dynamic Neural Relational Inference
Dong Gong, Frederic Z. Zhang, Javen Qinfeng Shi, Anton van den Hengel; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
 11843-11852
Dynamic interacting systems are prevalent in vision tasks. These interactions ar
e usually difficult to observe and measure directly, and yet understanding laten
t interactions is essential for performing inference tasks on dynamic systems li
ke forecasting. Neural relational inference (NRI) techniques are thus introduced
 to explicitly estimate interpretable relations between the entities in the syst
em for trajectory prediction. However, NRI assumes static relations; thus, dynam
ic neural relational inference (DNRI) was proposed to handle dynamic relations u
sing LSTM. Unfortunately, the older information will be washed away when the LST
M updates the latent variable as a whole, which is why DNRI struggles with model
ing long-term dependences and forecasting long sequences. This motivates us to p
ropose a memory-augmented dynamic neural relational inference method, which main
tains two associative memory pools: one for the interactive relations and the ot
her for the individual entities. The two memory pools help retain useful relatio
n features and node features for the estimation in the future steps. Our model d
ynamically estimates the relations by learning better embeddings and utilizing t
he long-range information stored in the memory. With the novel memory modules an
d customized structures, our memory-augmented DNRI can update and access the mem
ory adaptively as required. The memory pools also serve as global latent variabl
es across time to maintain detailed long-term temporal relations readily availab
le for other components to use. Experiments on synthetic and real-world datasets
 show the effectiveness of the proposed method on modeling dynamic relations and
 forecasting complex trajectories.
****

Physics-Based Differentiable Depth Sensor Simulation
Benjamin Planche, Rajat Vikram Singh; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2021, pp. 14387-14397
Gradient-based algorithms are crucial to modern computer-vision and graphics app
lications, enabling learning-based optimization and inverse problems. For exampl
e, photorealistic differentiable rendering pipelines for color images have been
proven highly valuable to applications aiming to map 2D and 3D domains. However,
 to the best of our knowledge, no effort has been made so far towards extending
these gradient-based methods to the generation of depth (2.5D) images, as simula
ting structured-light depth sensors implies solving complex light transport and
stereo-matching problems. In this paper, we introduce a novel end-to-end differe
ntiable simulation pipeline for the generation of realistic 2.5D scans, built on
 physics-based 3D rendering and custom block-matching algorithms. Each module ca
n be differentiated w.r.t sensor and scene parameters; e.g., to automatically tu
ne the simulation for new devices over some provided scans or to leverage the pi
peline as a 3D-to-2.5D transformer within larger computer-vision applications. A
pplied to the training of deep-learning methods for various depth-based recognit

ion tasks (classification, pose estimation, semantic segmentation), our simulati on greatly improves the performance of the resulting models on real scans, there by demonstrating the fidelity and value of its synthetic depth data compared to previous static simulations and learning-based domain adaptation schemes.
****

Temporal Action Detection With Multi-Level Supervision
Baifeng Shi, Qi Dai, Judy Hoffman, Kate Saenko, Trevor Darrell, Huijuan Xu; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021 , pp. 8022-8032
Training temporal action detection in videos requires large amounts of labeled d ata, yet such annotation is expensive to collect. Incorporating unlabeled or wea kly-labeled data to train action detection model could help reduce annotation co st. In this work, we first introduce the Semi-supervised Action Detection (SSAD) task with a mixture of labeled and unlabeled data and analyze different types o f errors in the proposed SSAD baselines which are directly adapted from the semi -supervised classification literature. Identifying that the main source of error is action incompleteness (i.e., missing parts of actions), we alleviate it by d esigning an unsupervised foreground attention (UFA) module utilizing the conditi onal independence between foreground and background motion. Then we incorporate weakly-labeled data into SSAD and propose Omni-supervised Action Detection (OSAD) with three levels of supervision. To overcome the accompanying action-context confusion problem in OSAD baselines, an information bottleneck (IB) is designed to suppress the scene information in non-action frames while preserving the acti on information. We extensively benchmark against the baselines for SSAD and OSAD on our created data splits in THUMOS14 and ActivityNet1.2, and demonstrate the effectiveness of the proposed UFA and IB methods. Lastly, the benefit of our ful l OSAD-IB model under limited annotation budgets is shown by exploring the optim al annotation strategy for labeled, unlabeled and weakly-labeled data.
****

FACIAL: Synthesizing Dynamic Talking Face With Implicit Attribute Learning
Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, Xiaohu Guo; Proceedings of the IEEE/CVF International Conference on Computer Vi sion (ICCV), 2021, pp. 3867-3876
In this paper, we propose a talking face generation method that takes an audio s ignal as input and a short target video clip as reference, and synthesizes a pho to-realistic video of the target face with natural lip motions, head poses, and eye blinks that are in-sync with the input audio signal. We note that the synthe tic face attributes include not only explicit ones such as lip motions that have high correlations with speech, but also implicit ones such as head poses and ey e blinks that have only weak correlation with the input audio. To model such com plicated relationships among different face attributes with input audio, we prop ose a FACe Implicit Attribute Learning Generative Adversarial Network (FACIAL-GA N), which integrates the phonetics-aware, context-aware, and identity-aware info rmation to synthesize the 3D face animation with realistic motions of lips, head poses, and eye blinks. Then, our Rendering-to-Video network takes the rendered face images and the attention map of eye blinks as input to generate the photo-r ealistic output video frames. Experimental results and user studies show our met hod can generate realistic talking face videos with not only synchronized lip mo tions, but also natural head movements and eye blinks, with better qualities tha n the results of state-of-the-art methods.
****

Unsupervised Deep Video Denoising
Dev Yashpal Sheth, Sreyas Mohan, Joshua L. Vincent, Ramon Manzorro, Peter A. Cro zier, Mitesh M. Khapra, Eero P. Simoncelli, Carlos Fernandez-Granda; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1 759-1768
Deep convolutional neural networks (CNNs) for video denoising are typically trai ned with supervision, assuming the availability of clean videos. However, in man y applications, such as microscopy, noiseless videos are not available. To addre ss this, we propose an Unsupervised Deep Video Denoiser (UDVD), a CNN architectu

re designed to be trained exclusively with noisy data. The performance of UDVD i s comparable to the supervised state-of-the-art, even when trained only on a sin gle short noisy video. We demonstrate the promise of our approach in real-world imaging applications by denoising raw video, fluorescence-microscopy and electro n-microscopy data. In contrast to many current approaches to video denoising, UD VD does not require explicit motion compensation. This is advantageous because m otion compensation is computationally expensive, and can be unreliable when the input data are noisy. A gradient-based analysis reveals that UDVD automatically tracks the motion of objects in the input noisy videos. Thus, the network learns to perform implicit motion compensation, even though it is only trained for den oising.

****

Making Higher Order MOT Scalable: An Efficient Approximate Solver for Lifted Dis joint Paths

Andrea Hornakova, Timo Kaiser, Paul Swoboda, Michal Rolinek, Bodo Rosenhahn, Rob erto Henschel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6330-6340

We present an efficient approximate message passing solver for the lifted disjoi nt paths problem (LDP), a natural but NP-hard model for multiple object tracking (MOT). Our tracker scales to very large instances that come from long and crowd ed MOT sequences. Our approximate solver enables us to process the MOT15/16/17 b enchmarks without sacrificing solution quality and allows for solving MOT20, whi ch has been out of reach up to now for LDP solvers due to its size and complexit y. On all these four standard MOT benchmarks we achieve performance comparable o r better than current state-of-the-art methods including a tracker based on an o ptimal LDP solver.

****

TMCOSS: Thresholded Multi-Criteria Online Subset Selection for Data-Efficient Au tonomous Driving

Soumi Das, Harikrishna Patibandla, Suparna Bhattacharya, Kshounis Bera, Niloy Ga nguly, Sourangshu Bhattacharya; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 6341-6350

Training vision-based Autonomous driving models is a challenging problem with en ormous practical implications. One of the main challenges is the requirement of storage and processing of vast volumes of (possibly redundant) driving video dat a. In this paper, we study the problem of data-efficient training of autonomous driving systems. We argue that in the context of an edge-device deployment, mult i-criteria online video frame subset selection is an appropriate technique for d eveloping such frameworks. We study existing convex optimization based solutions and show that they are unable to provide solution with high weightage to loss o f selected video frames. We design a novel multi-criteria online subset selectio n algorithm, TMCOSS, which uses a thresholded concave function of selection vari ables. Extensive experiments using driving simulator CARLA show that we are able to drop 80% of the frames, while succeeding to complete 100% of the episodes. W e also show that TMCOSS improves performance on the crucial affordance 'Relative Angle' during turns, on inclusion of bucket-specific relative angle loss (BL), leading to selection of more frames in those parts. TMCOSS also achieves an 80% reduction in number of training video frames, on real-world videos from the stan dard BDD and Cityscapes datasets, for the tasks of drivable area segmentation, a nd semantic segmentation.

****

Efficient Visual Pretraining With Contrastive Detection

Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Or iol Vinyals, João Carreira; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10086-10096

Self-supervised pretraining has been shown to yield powerful representations for transfer learning. These performance gains come at a large computational cost h owever, with state-of-the-art methods requiring an order of magnitude more compu tation than supervised pretraining. We tackle this computational bottleneck by i ntroducing a new self-supervised objective, contrastive detection, which tasks r

epresentations with identifying object-level features across augmentations. This objective extracts a rich learning signal per image, leading to state-of-the-art transfer accuracy on a variety of downstream tasks, while requiring up to 10x less pretraining. In particular, our strongest ImageNet-pretrained model performs on par with SEER, one of the largest self-supervised systems to date, which uses 1000x more pretraining data. Finally, our objective seamlessly handles pretraining on more complex images such as those in COCO, closing the gap with supervised transfer learning from COCO to PASCAL.

****

Exploiting Scene Graphs for Human-Object Interaction Detection

Tao He, Lianli Gao, Jingkuan Song, Yuan-Fang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15984-15993

Human-Object Interaction (HOI) detection is a fundamental visual task aiming at localizing and recognizing interactions between humans and objects. Existing works focus on the visual and linguistic features of humans and objects. However, they do not captalise on the high-level and semantic relationships present in the image, which provides crucial contextual and detailed relational knowledge for HOI inference. We propose a novel method to exploit this information, through the scene graph, for the HumanObject Interaction (SG2HOI) detection task. Our method, SG2HOI, incorporates the SG information in two ways: (1) we embed a scene graph into a global context clue, serving as the scene-specific environmental context; and (2) we build a relation-aware message-passing module to gather relationships from objects' neighborhood and transfer them into interactions. Empirical evaluation shows that our SG2HOI method outperforms the state-of-the-art methods on two benchmark HOI datasets: V-COCO and HICO-DET. Code will be available at https://github.com/ht014/SG2HOI.

****

Multi-VAE: Learning Disentangled View-Common and View-Peculiar Visual Representations for Multi-View Clustering

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, Lifang He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9234-9243

Multi-view clustering, a long-standing and important research problem, focuses on mining complementary information from diverse views. However, existing works often fuse multiple views' representations or handle clustering in a common feature space, which may result in their entanglement especially for visual representations. To address this issue, we present a novel VAE-based multi-view clustering framework (Multi-VAE) by learning disentangled visual representations. Concretely, we define a view-common variable and multiple view-peculiar variables in the generative model. The prior of view-common variable obeys approximately discrete Gumbel Softmax distribution, which is introduced to extract the common cluster factor of multiple views. Meanwhile, the prior of view-peculiar variable follows continuous Gaussian distribution, which is used to represent each view's peculiar visual factors. By controlling the mutual information capacity to disentangle the view-common and view-peculiar representations, continuous visual information of multiple views can be separated so that their common discrete cluster information can be effectively mined. Experimental results demonstrate that Multi-VAE enjoys the disentangled and explainable visual representations, while obtaining superior clustering performance compared with state-of-the-art methods.

****

TF-Blender: Temporal Feature Blender for Video Object Detection

Yiming Cui, Liqi Yan, Zhiwen Cao, Dongfang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8138-8147

Video objection detection is a challenging task because isolated video frames may encounter appearance deterioration, which introduces great confusion for detection. One of the popular solutions is to exploit the temporal information and enhance per-frame representation through aggregating features from neighboring frames. Despite achieving improvements in detection, existing methods focus on the selection of higher-level video frames for aggregation rather than modeling lower-level temporal relations to increase the feature representation. To address th

is limitation, we propose a novel solution named TF-Blender, which includes thre e modules: 1) Temporal relation models the relations between the current frame a nd its neighboring frames to preserve spatial information. 2). Feature adjustmen t enriches the representation of every neighboring feature map; 3) Feature blend er combines outputs from the first two modules and produces stronger features fo r the later detection tasks. For its simplicity, TF-Blender can be effortlessly plugged into any detection network to improve detection behavior. Extensive eval uations on ImageNet VID and YouTube-VIS benchmarks indicate the performance guar antees of using TF-Blender on recent state-of-the-art methods.
****

Adversarial Robustness for Unsupervised Domain Adaptation
Muhammad Awais, Fengwei Zhou, Hang Xu, Lanqing Hong, Ping Luo, Sung-Ho Bae, Zhen guo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8568-8577
Extensive Unsupervised Domain Adaptation (UDA) studies have shown great success in practice by learning transferable representations across a labeled source dom ain and an unlabeled target domain with deep models. However, current work focus es on improving the generalization ability of UDA models on clean examples witho ut considering the adversarial robustness, which is crucial in real-world applic ations. Conventional adversarial training methods are not suitable for the adver sarial robustness on the unlabeled target domain of UDA since they train models with adversarial examples generated by the supervised loss function. In this wor k, we propose to leverage intermediate representations learned by robust ImageNe t models to improve the robustness of UDA models. Our method works by aligning t he features of the UDA model with the robust features learned by ImageNet pre-tr ained models along with domain adaptation training. It utilizes both labeled and unlabeled domains and instills robustness without any adversarial intervention or label requirement during domain adaptation training. Our experimental results show that our method significantly improves adversarial robustness compared to the baseline while keeping clean accuracy on various UDA benchmarks.
****

Discovering 3D Parts From Image Collections
Chun-Han Yao, Wei-Chih Hung, Varun Jampani, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12981-129 90
Reasoning 3D shapes from 2D images is an essential yet challenging task, especia lly when only single-view images are at our disposal. While an object can have a complicated shape, individual parts are usually close to geometric primitives a nd thus are easier to model. Furthermore, parts provide a mid-level representati on that is robust to appearance variations across objects in a particular catego ry. In this work, we tackle the problem of 3D part discovery from only 2D image collections. Instead of relying on manually annotated parts for supervision, we propose a self-supervised approach, latent part discovery (LPD). Our key insight is to learn a novel part shape prior that allows each part to fit an object sha pe faithfully while constrained to have simple geometry. Extensive experiments o n the synthetic ShapeNet, PartNet, and real-world Pascal 3D+ datasets show that our method discovers consistent object parts and achieves favorable reconstructi on accuracy compared to the existing methods with the same level of supervision. Our project page with code is at https://chhankyao.github.io/lpd/.
****

ICE: Inter-Instance Contrastive Encoding for Unsupervised Person Re-Identificati on
Hao Chen, Benoit Lagadec, François Bremond; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2021, pp. 14960-14969
Unsupervised person re-identification (ReID) aims at learning discriminative ide ntity features without annotations. Recently, self-supervised contrastive learni ng has gained increasing attention for its effectiveness in unsupervised represe ntation learning. The main idea of instance contrastive learning is to match a s ame instance in different augmented views. However, the relationship between dif ferent instances has not been fully explored in previous contrastive methods, es

pecially for instance-level contrastive loss. To address this issue, we propose Inter-instance Contrastive Encoding (ICE) that leverages inter-instance pairwise similarity scores to boost previous class-level contrastive ReID methods. We first use pairwise similarity ranking as one-hot hard pseudo labels for hard instance contrast, which aims at reducing intra-class variance. Then, we use similarity scores as soft pseudo labels to enhance the consistency between augmented and original views, which makes our model more robust to augmentation perturbations. Experiments on several large-scale person ReID datasets validate the effectiveness of our proposed unsupervised method ICE, which is competitive with even supervised methods. Code is made available at https://github.com/chenhao2345/ICE.

****

PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, Shan Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13759-13768

Generating portrait images by controlling the motions of existing faces is an important task of great consequence to social media industries. For easy use and intuitive control, semantically meaningful and fully disentangled parameters should be used as modifications. However, many existing techniques do not provide such fine-grained controls or use indirect editing methods i.e. mimic motions of other individuals. In this paper, a Portrait Image Neural Renderer (PIRenderer) is proposed to control the face motions with the parameters of three-dimensional morphable face models (3DMMs). The proposed model can generate photo-realistic portrait images with accurate movements according to intuitive modifications. Experiments on both direct and indirect editing tasks demonstrate the superiority of this model. Meanwhile, we further extend this model to tackle the audio-driven facial reenactment task by extracting sequential motions from audio inputs. We show that our model can generate coherent videos with convincing movements from only a single reference image and a driving audio stream. Our source code is available at https://github.com/RenYurui/PIRender.

****

Toward Human-Like Grasp: Dexterous Grasping via Semantic Representation of Object-Hand

Tianqiang Zhu, Rina Wu, Xiangbo Lin, Yi Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15741-15751

In recent years, many dexterous robotic hands have been designed to assist or replace human hands in executing various tasks. But how to teach them to perform dexterous operations like human hands is still a challenging task. In this paper, we propose a grasp synthesis framework to make robots grasp and manipulate objects like human beings. We first build a dataset by accurately segmenting the functional areas of the object and annotating semantic touch code for each functional area to guide the dexterous hand to complete the functional grasp and post-grasp manipulation. This dataset contains 18 categories of 129 objects selected from four datasets, and 15 people participated in data annotation. Then we carefully design four loss functions to constrain the network, which successfully generates the functional grasp of dexterous hand under the guidance of semantic touch code. The thorough experiments in synthetic data show our model can robustly generate functional grasp, even for objects that the model has not see before.

****

MAAS: Multi-Modal Assignation for Active Speaker Detection

Juan Léon Alcázar, Fabian Caba, Ali K. Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 265-274

Active speaker detection requires a solid integration of multi-modal cues. While individual modalities can approximate a solution, accurate predictions can only be achieved by explicitly fusing the audio and visual features and modeling their temporal progression. Despite its inherent muti-modal nature, current methods still focus on modeling and fusing short-term audiovisual features for individual speakers, often at frame level. In this paper we present a novel approach to active speaker detection that directly addresses the multi-modal nature of the problem, and provides a straightforward strategy where independent visual features from potential speakers in the scene are assigned to a previously detected spe

ech event. Our experiments show that, an small graph data structure built from l
ocal information, allows to approximate an instantaneous audio-visual assignment
 problem. Moreover, the temporal extension of this initial graph achieves a new
state-of-the-art performance on the AVA-ActiveSpeaker dataset with a mAP of 88.8
%.
****

Multi-Source Domain Adaptation for Object Detection
Xingxu Yao, Sicheng Zhao, Pengfei Xu, Jufeng Yang; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 3273-3282
To reduce annotation labor associated with object detection, an increasing numbe
r of studies focus on transferring the learned knowledge from a labeled source d
omain to another unlabeled target domain. However, existing methods assume that
the labeled data are sampled from a single source domain, which ignores a more g
eneralized scenario, where labeled data are from multiple source domains. For th
e more challenging task, we propose a unified Faster RCNN based framework, terme
d Divide-and-Merge Spindle Network (DMSN), which can simultaneously enhance doma
in invariance and preserve discriminative power. Specifically, the framework con
tains multiple source subnets and a pseudo target subnet. First, we propose a hi
erarchical feature alignment strategy to conduct strong and weak alignments for
low- and high-level features, respectively, considering their different effects
for object detection. Second, we develop a novel pseudo subnet learning algorith
m to approximate optimal parameters of pseudo target subset by weighted combinat
ion of parameters in different source subnets. Finally, a consistency regulariza
tion for region proposal network is proposed to facilitate each subnet to learn
more abstract invariances. Extensive experiments on different adaptation scenari
os demonstrate the effectiveness of the proposed model.
****

Learning Conditional Knowledge Distillation for Degraded-Reference Image Quality
 Assessment
Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10242
-10251
An important scenario for image quality assessment (IQA) is to evaluate image re
storation (IR) algorithms. The state-of-the-art approaches adopt a full-referenc
e paradigm that compares restored images with their corresponding pristine-quali
ty images. However, pristine-quality images are usually unavailable in blind ima
ge restoration tasks and real-world scenarios. In this paper, we propose a pract
ical solution named degraded-reference IQA (DR-IQA), which exploits the inputs o
f IR models, degraded images, as references. Specifically, we extract reference
information from degraded images by distilling knowledge from pristine-quality i
mages. The distillation is achieved through learning a reference space, where va
rious degraded images are encouraged to share the same feature statistics with p
ristine-quality images. And the reference space is optimized to capture deep ima
ge priors that are useful for quality assessment. Note that pristine-quality ima
ges are only used during training. Our work provides a powerful and differentiab
le metric for blind IRs, especially for GAN-based methods. Extensive experiments
 show that our results can even be close to the performance of full-reference se
ttings.
****

ShapeConv: Shape-Aware Convolutional Layer for Indoor RGB-D Semantic Segmentatio
n
Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, Yangyan
 Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2021, pp. 7088-7097
RGB-D semantic segmentation has attracted increasing attention over the past few
 years. Existing methods mostly employ homogeneous convolution operators to cons
ume the RGB and depth features, ignoring their intrinsic differences. In fact, t
he RGB values capture the photometric appearance properties in the projected ima
ge space, while the depth feature encodes both the shape of a local geometry as
well as the base (whereabout) of it in a larger context. Compared with the base,

the shape probably is more inherent and has a stronger connection to the semantics, and thus is more critical for segmentation accuracy. Inspired by this observation, we introduce Shape-aware Convolutional layer (ShapeConv) for processing the depth feature, where the depth feature is firstly decomposed into a shape-component and a base-component, next two learnable weights are introduced to cooperate with them independently, and finally a convolution is applied on the re-weighted combination of these two components. ShapeConv is model-agnostic and can be easily integrated into most CNNs to replace vanilla convolutional layers for semantic segmentation. Extensive experiments on three challenging indoor RGB-D semantic segmentation benchmarks, i.e., NYU-Dv2(-13,-40), SUN RGB-D, and SID, demonstrate the effectiveness of our ShapeConv when employing it over five popular architectures. Moreover, the performance of CNNs with ShapeConv is boosted without introducing any computation and memory increase in the inference phase. The reason is that the learnt weights for balancing the importance between the shape and base components in ShapeConv become constants in the inference phase, and thus can be fused into the following convolution, resulting in a network that is identical to one with vanilla convolutional layers.

****

GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, Serena Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3942-3951

In recent years, the growing number of medical imaging studies is placing an ever-increasing burden on radiologists. Deep learning provides a promising solution for automatic medical image analysis and clinical decision support. However, large-scale manually labeled datasets required for training deep neural networks are difficult and expensive to obtain for medical images. The purpose of this work is to develop label-efficient multimodal medical imaging representations by leveraging radiology reports. Specifically, we propose an attention-based framework (GLoRIA) for learning global and local representations by contrasting image sub-regions and words in the paired report. In addition, we propose methods to leverage the learned representations for various downstream medical image recognition tasks with limited labels. Our results demonstrate high-performance and label-efficiency for image-text retrieval, classification (finetuning and zeros-shot settings), and segmentation on different datasets.

****

Summarize and Search: Learning Consensus-Aware Dynamic Convolution for Co-Saliency Detection

Ni Zhang, Junwei Han, Nian Liu, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4167-4176

Humans perform co-saliency detection by first summarizing the consensus knowledge in the whole group and then searching corresponding objects in each image. Previous methods usually lack robustness, scalability, or stability for the first process and simply fuse consensus features with image features for the second process. In this paper, we propose a novel consensus-aware dynamic convolution model to explicitly and effectively perform the "summarize and search" process. To summarize consensus image features, we first summarize robust features for every single image using an effective pooling method and then aggregate cross-image consensus cues via the self-attention mechanism. By doing this, our model meets the scalability and stability requirements. Next, we generate dynamic kernels from consensus features to encode the summarized consensus knowledge. Two kinds of kernels are generated in a supplementary way to summarize fine-grained image-specific consensus object cues and the coarse group-wise common knowledge, respectively. Then, we can effectively perform object searching by employing dynamic convolution at multiple scales. Besides, a novel and effective data synthesis method is also proposed to train our network. Experimental results on four benchmark datasets verify the effectiveness of our proposed method. Our code and saliency maps are available at https://github.com/nnizhang/CADC.

****

Visual Distant Supervision for Scene Graph Generation
Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Werm ter, Maosong Sun; Proceedings of the IEEE/CVF International Conference on Comput er Vision (ICCV), 2021, pp. 15816-15826
Scene graph generation aims to identify objects and their relations in images, p roviding structured image representations that can facilitate numerous applicati ons in computer vision. However, scene graph models usually require supervised l earning on large quantities of labeled data with intensive human annotation. In this work, we propose visual distant supervision, a novel paradigm of visual rel ation learning, which can train scene graph models without any human-labeled dat a. The intuition is that by aligning commonsense knowledge bases and images, we can automatically create large-scale labeled data to provide distant supervision for visual relation learning. To alleviate the noise in distantly labeled data, we further propose a framework that iteratively estimates the probabilistic rel ation labels and eliminates the noisy ones. Comprehensive experimental results s how that our distantly supervised model outperforms strong weakly supervised and semi-supervised baselines. By further incorporating human-labeled data in a sem i-supervised fashion, our model outperforms state-of-the-art fully supervised mo dels by a large margin (e.g., 8.3 micro- and 7.8 macro-recall@50 improvements fo r predicate classification in Visual Genome evaluation). We make the data and co de for this paper publicly available at https://github.com/thunlp/VisualDS.
****

Viewpoint-Agnostic Change Captioning With Cycle Consistency
Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, Gunhee Kim; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2095-2104
Change captioning is the task of identifying the change and describing it with a concise caption. Despite recent advancements, filtering out insignificant chang es still remains as a challenge. Namely, images from different camera perspectiv es can cause issues; a mere change in viewpoint should be disregarded while stil l capturing the actual changes. In order to tackle this problem, we present a ne w Viewpoint-Agnostic change captioning network with Cycle Consistency (VACC) tha t requires only one image each for the before and after scene, without depending on any other information. We achieve this by devising a new difference encoder module which can encode viewpoint information and model the difference more effe ctively. In addition, we propose a cycle consistency module that can potentially improve the performance of any change captioning networks in general by matchin g the composite feature of the generated caption and before image with the after image feature. We evaluate the performance of our proposed model across three d atasets for change captioning, including a novel dataset we introduce here that contains images with changes under extreme viewpoint shifts. Through our experim ents, we show the excellence of our method with respect to the CIDEr, BLEU-4, ME TEOR and SPICE scores. Moreover, we demonstrate that attaching our proposed cycl e consistency module yields a performance boost for existing change captioning n etworks, even with varying image encoding mechanisms.
****

Neural Video Portrait Relighting in Real-Time via Consistency Modeling
Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, Lan Xu; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 802-812
Video portraits relighting is critical in user-facing human photography, especia lly for immersive VR/AR experience. Recent advances still fail to recover consis tent relit result under dynamic illuminations from monocular RGB stream, sufferi ng from the lack of video consistency supervision. In this paper, we propose a n eural approach for real-time, high-quality and coherent video portrait relightin g, which jointly models the semantic, temporal and lighting consistency using a new dynamic OLAT dataset. We propose a hybrid structure and lighting disentangle ment in an encoder-decoder architecture, which combines a multi-task and adversa rial training strategy for semantic-aware consistency modeling. We adopt a tempo ral modeling scheme via flow-based supervision to encode the conjugated temporal consistency in a cross manner. We also propose a lighting sampling strategy to

model the illumination consistency and mutation for natural portrait light manip
ulation in real-world. Extensive experiments demonstrate the effectiveness of ou
r approach for consistent video portrait light-editing and relighting, even usin
g mobile computing.
****

Image Shape Manipulation From a Single Augmented Training Sample
Yael Vinker, Eliahu Horwitz, Nir Zabari, Yedid Hoshen; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2021, pp. 13769-13778
In this paper, we present DeepSIM, a generative model for conditional image mani
pulation based on a single image. We find that extensive augmentation is key for
 enabling single image training, and incorporate the use of thin-plate-spline (T
PS) as an effective augmentation. Our network learns to map between a primitive
representation of the image to the image itself. The choice of a primitive repre
sentation has an impact on the ease and expressiveness of the manipulations and
can be automatic (e.g. edges), manual (e.g. segmentation) or hybrid such as edge
s on top of segmentations. At manipulation time, our generator allows for making
 complex image changes by modifying the primitive input representation and mappi
ng it through the network. Our method is shown to achieve remarkable performance
 on image manipulation tasks.
****

SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit S
hapes
Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, Andreas Geiger; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 11594-11604
Neural implicit surface representations have emerged as a promising paradigm to
capture 3D shapes in a continuous and resolution-independent manner. However, ad
apting them to articulated shapes is non-trivial. Existing approaches learn a ba
ckward warp field that maps deformed to canonical points. However, this is probl
ematic since the backward warp field is pose dependent and thus requires large a
mounts of data to learn. To address this, we introduce SNARF, which combines the
 advantages of linear blend skinning (LBS) for polygonal meshes with those of ne
ural implicit surfaces by learning a forward deformation field without direct su
pervision. This deformation field is defined in canonical, pose-independent, spa
ce, enabling generalization to unseen poses. Learning the deformation field from
 posed meshes alone is challenging since the correspondences of deformed points
are defined implicitly and may not be unique under changes of topology. We propo
se a forward skinning model that finds all canonical correspondences of any defo
rmed point using iterative root finding. We derive analytical gradients via impl
icit differentiation, enabling end-to-end training from 3D meshes with bone tran
sformations. Compared to state-of-the-art neural implicit representations, our a
pproach generalizes better to unseen poses while preserving accuracy. We demonst
rate our method in challenging scenarios on (clothed) 3D humans in diverse and u
nseen poses.
****

Curvature Generation in Curved Spaces for Few-Shot Learning
Zhi Gao, Yuwei Wu, Yunde Jia, Mehrtash Harandi; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2021, pp. 8691-8700
Few-shot learning describes the challenging problem of recognizing samples from
unseen classes given very few labeled examples. In many cases, few-shot learning
 is cast as learning an embedding space that assigns test samples to their corre
sponding class prototypes. Previous methods assume that data of all few-shot lea
rning tasks comply with a fixed geometrical structure, mostly a Euclidean struct
ure. Questioning this assumption that is clearly difficult to hold in real-world
 scenarios and incurs distortions to data, we propose to learn a task-aware curv
ed embedding space by making use of the hyperbolic geometry. As a result, task-s
pecific embedding spaces where suitable curvatures are generated to match the ch
aracteristics of data are constructed, leading to more generic embedding spaces.
 We then leverage on intra-class and inter-class context information in the embe
dding space to generate class prototypes for discriminative classification. We c

onduct a comprehensive set of experiments on inductive and transductive few-shot learning, demonstrating the benefits of our proposed method over existing embedding methods.

****

Single Image 3D Shape Retrieval via Cross-Modal Instance and Category Contrastive Learning

Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, Lin Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11405-11415

In this work, we tackle the problem of single image-based 3D shape retrieval (IBSR), where we seek to find the most matched shape of a given single 2D image from a shape repository. Most of the existing works learn to embed 2D images and 3D shapes into a common feature space and perform metric learning using a triplet loss. Inspired by the great success in recent contrastive learning works on self-supervised representation learning, we propose a novel IBSR pipeline leveraging contrastive learning. We note that adopting such cross-modal contrastive learning between 2D images and 3D shapes into IBSR tasks is non-trivial and challenging: contrastive learning requires very strong data augmentation in constructed positive pairs to learn the feature invariance, whereas traditional metric learning works do not have this requirement. Moreover, object shape and appearance are entangled in 2D query images, thus making the learning task more difficult than contrasting single-modal data. To mitigate the challenges, we propose to use multi-view grayscale rendered images from the 3D shapes as a shape representation. We then introduce a strong data augmentation technique based on color transfer, which can significantly but naturally change the appearance of the query image, effectively satisfying the need for contrastive learning. Finally, we propose to incorporate a novel category-level contrastive loss that helps distinguish similar objects from different categories, in addition to classic instance-level contrastive loss. Our experiments demonstrate that our approach achieves the best performance on all the three popular IBSR benchmarks, including Pix3D, Stanford Cars, and Comp Cars, outperforming the previous state-of-the-art from 4% - 15% on retrieval accuracy.

****

Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, Amir Zamir; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10786-10796

Computer vision now relies on data, but we know surprisingly little about what factors in the data affect performance. We argue that this stems from the way data is collected. Designing and collecting static datasets of images (or videos) locks us in to specific design choices and limits us to post-hoc analyses. In practice, vision datasets only include specific domains and tasks. This not only makes it necessary and difficult to combine datsets, but leads to scattershot overall coverage that frustrates systemic research into the interaction of tasks, data, models, and learning algorithms. For example, if a model trained for ImageNet classification on ImageNet transfers better to CoCo than does a model trained for Kitti depth estimation--is that due to the difference in tasks or the different training data? We note that one way to do this is to use a comprehensive, standardized scene representation that contains extra information about the scene, and then to use that to create a specific dataset of study. We introduce a platform for doing this. Specifically, we provide a pipeline that takes as input a 3D scans and generates multi-task datasets of mid-level cues. The pipeline exposes complete control over the generation process, is implemented in mostly python, and we provide ecosystem tools such as a Docker and PyTorch dataloaders. We also provide a starter dataset of several recent 3D scan datasets, processed into standard static datasets of mid-level cues. We show that this starter dataset (generated from the annotator pipeline) is reliable; it yields models that provide state-of-the-art performance for several tasks. It yields human-level surface normal estimation performance on OASIS, despite having never seen OASIS data during training. With the proliferation of cheaper 3D sensors (e.g. on the newest iPh

one), we anticipate that releasing an automated tool for this processing pipelin e will allow the starter set to continue to expand and cover more domains. We ex amine a few small examples of using this procedure to analyze the relationship o f data, tasks, models and learning algorithms, and suggest several exciting dire ctions that are well out of the scope of this paper.
****

Single View Physical Distance Estimation Using Human Pose
Xiaohan Fei, Henry Wang, Lin Lee Cheong, Xiangyu Zeng, Meng Wang, Joseph Tighe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12406-12416
We propose a fully automated system that simultaneously estimates the camera int rinsics, the ground plane, and physical distances between people from a single R GB image or video captured by a camera viewing a 3-D scene from a fixed vantage point. To automate camera calibration and distance estimation, we leverage prior s about human pose and develop a novel direct formulation for pose-based auto-ca libration and distance estimation, which shows state-of-the-art performance on p ublicly available datasets. The proposed approach enables existing camera system s to measure physical distances without needing a dedicated calibration process or range sensors, and is applicable to a broad range of use cases such as social distancing and workplace safety. Furthermore, to enable evaluation and drive re search in this area, we contribute to the publicly available MEVA dataset with a dditional distance annotations, resulting in "MEVADA" -- an evaluation benchmark for the pose-based auto-calibration and distance estimation problem.
****

Few-Shot Semantic Segmentation With Cyclic Memory Network
Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, Ling Shao; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 7293-7302
Few-shot semantic segmentation (FSS) is an important task for novel (unseen) obj ect segmentation under the data-scarcity scenario. However, most FSS methods rel y on unidirectional feature aggregation, e.g., from support prototypes to get th e query prediction, and from high-resolution features to guide the low-resolutio n ones. This usually fails to fully capture the cross-resolution feature relatio nships and thus leads to inaccurate estimates of the query objects. To resolve t he above dilemma, we propose a cyclic memory network (CMN) to directly learn to read abundant support information from all resolution features in a cyclic manne r. Specifically, we first generate N pairs (key and value) of multi-resolution q uery features guided by the support feature and its mask. Next, we circularly ta ke one pair of these features as the query to be segmented, and the rest N-1 pai rs are written into an external memory accordingly, i.e., this leave-one-out pro cess is conducted for N times. In each cycle, the query feature is updated by co llaboratively matching its key and value with the memory, which can elegantly co ver all the spatial locations from different resolutions. Furthermore, we incorp orate the query feature re-adding and the query feature recursive updating mecha nisms into the memory reading operation. CMN, equipped with these merits, can th us capture cross-resolution relationships and better handle the object appearanc e and scale variations in FSS. Experiments on PASCAL-5i and COCO-20i well valida te the effectiveness of our model for FSS.
****

Weakly-Supervised Action Segmentation and Alignment via Transcript-Aware Union-o f-Subspaces Learning
Zijia Lu, Ehsan Elhamifar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8085-8095
We address the problem of learning to segment actions from weakly-annotated vide os, i.e., videos accompanied by transcripts (ordered list of actions). We propos e a framework in which we model actions with a union of low-dimensional subspace s, learn the subspaces using transcripts and refine video features that lend the mselves to action subspaces. To do so, we design an architecture consisting of a Union-of-Subspace Network, which is an ensemble of autoencoders, each modeling a low-dimensional action subspace and can capture variations of an action within and across videos. For learning, at each iteration, we generate positive and ne

gative soft alignment matrices using the segmentations from the previous iterati on, which we use for discriminative training of our model. To regularize the lea rning, we introduce a constraint loss that prevents imbalanced segmentations and enforces relatively similar duration of each action across videos. To have a re al-time inference, we develop a hierarchical segmentation framework that uses su bset selection to find representative transcripts and hierarchically align a tes t video with increasingly refined representative transcripts. Our experiments on three datasets show that our method improves the state-of-the-art action segmen tation and alignment, while speeding up the inference time by a factor of 4 to 1 3.

****

Not All Operations Contribute Equally: Hierarchical Operation-Adaptive Predictor for Neural Architecture Search

Ziye Chen, Yibing Zhan, Baosheng Yu, Mingming Gong, Bo Du; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10508-10517

Graph-based predictors have recently shown promising results on neural architect ure search (NAS). Despite their efficiency, current graph-based predictors treat all operations equally, resulting in biased topological knowledge of cell archi tectures. Intuitively, not all operations are equally significant during forward ing propagation when aggregating information from these operations to another op eration. To address the above issue, we propose a Hierarchical Operation-adaptiv e Predictor (HOP) for NAS. HOP contains an operation-adaptive attention module (OAM) to capture the diverse knowledge between operations by learning the relativ e significance of operations in cell architectures during aggregation over itera tions. In addition, a cell-hierarchical gated module (CGM) further refines and e nriches the obtained topological knowledge of cell architectures, by integrating cell information from each iteration of OAM. The experimental results compared with state-of-the-art predictors demonstrate the capability of our proposed HOP. In specific, only using 0.1% training data, HOP improves kendall's Tau by 3.45% , N@5 by 20 places on NASBech-101; only using 1% training data, HOP improves ken dall's Tau by 2.12%, N@5 by 18 places on NASBench-201, respectively.

****

SOTR: Segmenting Objects With Transformers

Ruohao Guo, Dantong Niu, Liao Qu, Zhenbo Li; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 7157-7166

Most recent transformer-based models show impressive performance on vision tasks , even better than Convolution Neural Networks (CNN). In this work, we present a novel, flexible, and effective transformer-based model for high-quality instanc e segmentation. The proposed method, Segmenting Objects with TRansformers (SOTR) , simplifies the segmentation pipeline, building on an alternative CNN backbone appended with two parallel subtasks: (1) predicting per-instance category via tr ansformer and (2) dynamically generating segmentation mask with the multi-level upsampling module. SOTR can effectively extract lower-level feature representati ons and capture long-range context dependencies by Feature Pyramid Network (FPN) and twin transformer, respectively. Meanwhile, compared with the original trans former, the proposed twin transformer is timeand resource-efficient since only a row and a column attention are involved to encode pixels. Moreover, SOTR is eas y to be incorporated with various CNN backbones and transformer model variants t o make considerable improvements for the segmentation accuracy and training conv ergence. Extensive experiments show that our SOTR performs well on the MS COCO d ataset and surpasses state-of-the-art instance segmentation approaches. We hope our simple but strong framework could serve as a preferment baseline for instanc e-level recognition. Our code is available at https://github.com/easton-cau/SOTR .

****

Adaptive Surface Normal Constraint for Depth Estimation

Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12849-12858

We present a novel method for single image depth estimation using surface normal

constraints. Existing depth estimation methods either suffer from the lack of g
eometric constraints, or are limited to the difficulty of reliably capturing geo
metric context, which leads to a bottleneck of depth estimation quality. We ther
efore introduce a simple yet effective method, named Adaptive Surface Normal (AS
N) constraint, to effectively correlate the depth estimation with geometric cons
istency. Our key idea is to adaptively determine the reliable local geometry fro
m a set of randomly sampled candidates to derive surface normal constraint, for
which we measure the consistency of the geometric contextual features. As a resu
lt, our method can faithfully reconstruct the 3D geometry and is robust to local
shape variations, such as boundaries, sharp corners and noises. We conduct exte
nsive evaluations and comparisons using public datasets. The experimental result
s demonstrate our method outperforms the state-of-the-art methods and has superi
or efficiency and robustness.
****

Enriching Local and Global Contexts for Temporal Action Localization
Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, Gang Hua; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2021, pp. 13516-13525
Effectively tackling the problem of temporal action localization (TAL) necessita
tes a visual representation that jointly pursues two confounding goals, i.e., fi
ne-grained discrimination for temporal localization and sufficient visual invari
ance for action classification. We address this challenge by enriching both the
local and global contexts in the popular two-stage temporal localization framewo
rk, where action proposals are first generated followed by action classification
and temporal boundary regression. Our proposed model, dubbed ContextLoc, can be
divided into three sub-networks: L-Net, G-Net and P-Net. L-Net enriches the loc
al context via fine-grained modeling of snippet-level features, which is formula
ted as a query-and-retrieval process. G-Net enriches the global context via high
er-level modeling of the video-level representation. In addition, we introduce a
novel context adaptation module to adapt the global context to different propos
als. P-Net further models the context-aware inter-proposal relations. We explore
two existing models to be the P-Net in our experiments. The efficacy of our pro
posed method is validated by experimental results on the THUMOS14 (54.3% at tIoU
@0.5) and ActivityNet v1.3 (56.01% at tIoU@0.5) datasets, which outperforms rece
nt states of the art. Code is available at https://github.com/buxiangzhiren/Cont
extLoc.
****

Hypergraph Neural Networks for Hypergraph Matching
Xiaowei Liao, Yong Xu, Haibin Ling; Proceedings of the IEEE/CVF International Co
nference on Computer Vision (ICCV), 2021, pp. 1266-1275
Hypergraph matching is a useful tool to find feature correspondence by consideri
ng higher-order structural information. Recently, the employment of deep learnin
g has made great progress in the matching of graphs, suggesting its potential fo
r hypergraphs. Hence, in this paper, we present the first, to our best knowledge
, unified hypergraph neural network (HNN) solution for hypergraph matching. Spec
ifically, given two hypergraphs to be matched, we first construct an association
hypergraph over them and convert the hypergraph matching problem into a node cl
assification problem on the association hypergraph. Then, we design a novel hype
rgraph neural network to effectively solve the node classification problem. Bein
g end-to-end trainable, our proposed method, named HNN-HM, jointly learns all it
s components with improved optimization. For evaluation, HNN-HM is tested on var
ious benchmarks and shows a clear advantage over state-of-the-arts.
****

DRAEM - A Discriminatively Trained Reconstruction Embedding for Surface Anomaly
Detection
Vitjan Zavrtanik, Matej Kristan, Danijel Sko■aj; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 8330-8339
Visual surface anomaly detection aims to detect local image regions that signifi
cantly deviate from normal appearance. Recent surface anomaly detection methods
rely on generative models to accurately reconstruct the normal areas and to fail
on anomalies. These methods are trained only on anomaly-free images, and often

require hand-crafted post-processing steps to localize the anomalies, which proh
ibits optimizing the feature extraction for maximal detection capability. In add
ition to reconstructive approach, we cast surface anomaly detection primarily as
 a discriminative problem and propose a discriminatively trained reconstruction
anomaly embedding model (DRAEM). The proposed method learns a joint representati
on of an anomalous image and its anomaly-free reconstruction, while simultaneous
ly learning a decision boundary between normal and anomalous examples. The metho
d enables direct anomaly localization without the need for additional complicate
d post-processing of the network output and can be trained using simple and gene
ral anomaly simulations. On the challenging MVTec anomaly detection dataset, DRA
EM outperforms the current state-of-the-art unsupervised methods by a large marg
in and even delivers detection performance close to the fully-supervised methods
 on the widely used DAGM surface-defect detection dataset, while substantially o
utperforming them in localization accuracy.
****

Gaussian Fusion: Accurate 3D Reconstruction via Geometry-Guided Displacement Int
erpolation
Duo Chen, Zixin Tang, Zhenyu Xu, Yunan Zheng, Yiguang Liu; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5916-5925
Reconstructing delicate geometric details with consumer RGB-D sensors is challen
ging due to sensor depth and poses uncertainties. To tackle this problem, we pro
pose a unique geometry-guided fusion framework: 1) First, we characterize fusion
 correspondences with the geodesic curves derived from the mass transport proble
m, also known as the Monge-Kantorovich problem. Compared with the depth map back
-projection methods, the geodesic curves reveal the geometric structures of the
local surface. 2) Moving the points along the geodesic curves is the core of our
 fusion approach, guided by local geometric properties, i.e., Gaussian curvature
 and mean curvature. Compared with the state-of-the-art methods, our novel geome
try-guided displacement interpolation fully utilizes the meaningful geometric fe
atures of the local surface. It makes the reconstruction accuracy and completene
ss improved. Finally, a significant number of experimental results on real objec
t data verify the superior performance of the proposed method. Our technique ach
ieves the most delicate geometric details on thin objects for which the original
 depth map back-projection fusion scheme suffers from severe artifacts (See Fig.
1).
****

Frequency-Aware Spatiotemporal Transformers for Video Inpainting Detection
Bingyao Yu, Wanhua Li, Xiu Li, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 8188-8197
In this paper, we propose a frequency-aware spatiotemporal transformers for deep
 In this paper, we propose a Frequency-Aware Spatiotemporal Transformer (FAST) f
or video inpainting detection, which aims to simultaneously mine the traces of v
ideo inpainting from spatial, temporal, and frequency domains. Unlike existing d
eep video inpainting detection methods that usually rely on hand-designed attent
ion modules and memory mechanism, the proposed FAST have innate global self-atte
ntion mechanisms to capture the long-range relations. While existing video inpai
nting methods usually explore the spatial and temporal connections in a video, o
ur method employs a spatiotemporal transformer framework to detect the spatial c
onnections between patches and temporal dependency between frames. As the inpain
ted videos usually lack high frequency details, the proposed FAST simultaneously
 exploits the frequency domain information with a specifically designed decoder.
 Extensive experimental results demonstrate that our approach achieves very comp
etitive performance and generalizes well.
****

Virtual Multi-Modality Self-Supervised Foreground Matting for Human-Object Inter
action
Bo Xu, Han Huang, Cheng Lu, Ziwen Li, Yandong Guo; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 438-447
Most existing human matting algorithms tried to separate pure human-only foregro
und from the background. In this paper, we propose a Virtual Multi-modality Fore

ground Matting (VMFM) method to learn human-object interactive foreground (human and objects interacted with him or her) from a raw RGB image. The VMFM method requires no additional inputs, e.g. trimap or known background. We reformulate foreground matting as a self-supervised multi-modality problem: factor each input image into estimated depth map, segmentation mask, and interaction heatmap using three auto-encoders. In order to fully utilize the characteristics of each modality, we first train a dual encoder-to-decoder network to estimate the same alpha matte. Then we introduce a self-supervised method: Complementary Learning(CL) to predict deviation probability map and exchange reliable gradients across modalities without label. We conducted extensive experiments to analyze the effectiveness of each modality and the significance of different components in complementary learning. We demonstrate that our model outperforms the state-of-the-art methods.

****

Mutual Supervision for Dense Object Detection

Ziteng Gao, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3641-3650

The classification and regression head are both indispensable components to build up a dense object detector, which are usually supervised by the same training samples and thus expected to have consistency with each other for detecting objects accurately in final detection pipelines. In this paper, we break the convention of the same training samples for these two heads in dense detectors and explore a novel supervisory paradigm, termed as Mutual Supervision (MuSu), to respectively and mutually assign training samples for the classification and regression head to ensure this consistency. MuSu defines training samples for the regression head mainly based on classification predicting scores and in turn, defines samples for the classification head based on localization scores from the regression head. Experimental results show that the convergence of detectors trained by this mutual supervision is guaranteed and the effectiveness of the proposed method is verified on the challenging MS COCO benchmark. We also find that tiling more anchors at the same location benefits detectors and leads to further improvements under this training scheme. We hope this work can inspire further researches on the interaction of the classification and regression task in detection and the supervision paradigm for detectors, especially separately for these two heads.

****

Orthographic-Perspective Epipolar Geometry

Viktor Larsson, Marc Pollefeys, Magnus Oskarsson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5570-5578

In this paper we consider the epipolar geometry between orthographic and perspective cameras. We generalize many of the classical results for the perspective essential matrix to this setting and derive novel minimal solvers, not only for the calibrated case, but also for partially calibrated and non-central camera setups. While orthographic cameras might seem exotic, they occur naturally in many applications. They can e.g. model 2D maps (such as floor plans), aerial/satellite photography and even approximate narrow field-of-view cameras (e.g. from telephoto lenses). In our experiments we highlight various applications of the developed theory and solvers, including Radar-Camera calibration and aligning Structure-from-Motion models to aerial or satellite images.

****

Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, Dragomir Anguelov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9710-9719

As autonomous driving systems mature, motion forecasting has received increasing attention as a critical requirement for planning. Of particular importance are interactive situations such as merges, unprotected turns, etc., where predicting

individual object motion is not sufficient. Joint predictions of multiple objects are required for effective route planning. There has been a critical need for highquality motion data that is rich in both interactions and annotation to develop motion planning models. In this work, we introduce the most diverse interactive motion dataset to our knowledge, and provide specific labels for interacting objects suitable for developing joint prediction models. With over 100,000 scenes, each 20 seconds long at 10 Hz, our new dataset contains more than 570 hours of unique data over 1750 km of roadways. It was collected by mining for interesting interactions between vehicles, pedestrians, and cyclists across six cities within the United States. We use a high-accuracy 3D auto-labeling system to generate high quality 3D bounding boxes for each road agent, and provide corresponding high definition 3D maps for each scene. Furthermore, we introduce a new set of metrics that provides a comprehensive evaluation of both single agent and joint agent interaction motion forecasting models. Finally, we provide strong baseline models for individual agent prediction and joint-prediction. We hope that this new large-scale interactive motion dataset will provide new opportunities for advancing motion forecasting models.

****

Seminar Learning for Click-Level Weakly Supervised Semantic Segmentation
Hongjun Chen, Jinbao Wang, Hong Cai Chen, Xiantong Zhen, Feng Zheng, Rongrong Ji, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6920-6929

Annotation burden has become one of the biggest barriers to semantic segmentation. Approaches based on click-level annotations have therefore attracted increasing attention due to their superior trade-off between supervision and annotation cost. In this paper, we propose seminar learning, a new learning paradigm for semantic segmentation with click-level supervision. The fundamental rationale of seminar learning is to leverage the knowledge from different networks to compensate for insufficient information provided in click-level annotations. Mimicking a seminar, our seminar learning involves a teacher-student and a student-student module, where a student can learn from both skillful teachers and other students. The teacher-student module uses a teacher network based on the exponential moving average to guide the training of the student network. In the student-student module, heterogeneous pseudo-labels are proposed to bridge the transfer of knowledge among students to enhance each other's performance. Experimental results demonstrate the effectiveness of seminar learning, which achieves the new state-of-the-art performance of 72.51% (mIOU), surpassing previous methods by a large margin of up to 16.88% on the Pascal VOC 2012 dataset.

****

Retrieve in Style: Unsupervised Facial Feature Transfer and Retrieval
Min Jin Chong, Wen-Sheng Chu, Abhishek Kumar, David Forsyth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3887-3896

We present Retrieve in Style (RIS), an unsupervised framework for facial feature transfer and retrieval on real images. Recent work shows capabilities of transferring local facial features by capitalizing on the disentanglement property of the StyleGAN latent space. RIS improves existing art on the following: 1) Introducing more effective feature disentanglement to allow for challenging transfers (i.e., hair, pose) that were not shown possible in SoTA methods. 2) Eliminating the need for per-image hyperparameter tuning, and for computing a catalog over a large batch of images. 3) Enabling fine-grained face retrieval using disentangled facial features (e.g., eyes). To our best knowledge, this is the first work to retrieve face images at this fine level. 4) Demonstrating robust, natural editing on real images. Our qualitative and quantitative analyses show RIS achieves both high-fidelity feature transfers and accurate fine-grained retrievals on real images. We also discuss the responsible applications of RIS. Our code is available at https://github.com/mchong6/RetrieveInStyle.

****

Rethinking and Improving Relative Position Encoding for Vision Transformer
Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, Hongyang Chao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10033-1

0041
Relative position encoding (RPE) is important for transformer to capture sequence ordering of input tokens. General efficacy has been proven in natural language processing. However, in computer vision, its efficacy is not well studied and even remains controversial, e.g., whether relative position encoding can work equally well as absolute position? In order to clarify this, we first review existing relative position encoding methods and analyze their pros and cons when applied in vision transformers. We then propose new relative position encoding methods dedicated to 2D images, called image RPE (iRPE). Our methods consider directional relative distance modeling as well as the interactions between queries and relative position embeddings in self-attention mechanism. The proposed iRPE methods are simple and lightweight. They can be easily plugged into transformer blocks. Experiments demonstrate that solely due to the proposed encoding methods, DeiT and DETR obtain up to 1.5% (top-1 Acc) and 1.3% (mAP) stable improvements over their original versions on ImageNet and COCO respectively, without tuning any extra hyperparameters such as learning rate and weight decay. Our ablation and analysis also yield interesting findings, some of which run counter to previous understanding. Code and models are open-sourced at https://github.com/microsoft/Cream/tree/main/iRPE.
****

Meta-Aggregator: Learning To Aggregate for 1-Bit Graph Neural Networks
Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5301-5310
In this paper, we study a novel meta aggregation scheme towards binarizing graph neural networks (GNNs). We begin by developing a vanilla 1-bit GNN framework that binarizes both the GNN parameters and the graph features. Despite the lightweight architecture, we observed that this vanilla framework suffered from insufficient discriminative power in distinguishing graph topologies, leading to a dramatic drop in performance. This discovery motivates us to devise meta aggregators to improve the expressive power of vanilla binarized GNNs, of which the aggregation schemes can be adaptively changed in a learnable manner based on the binarized features. Towards this end, we propose two dedicated forms of meta neighborhood aggregators, an exclusive meta aggregator termed as Greedy Gumbel Neighborhood Aggregator (GNA), and a diffused meta aggregator termed as Adaptable Hybrid Neighborhood Aggregator (ANA). GNA learns to exclusively pick one single optimal aggregator from a pool of candidates, while ANA learns a hybrid aggregation behavior to simultaneously retain the benefits of several individual aggregators. Furthermore, the proposed meta aggregators may readily serve as a generic plugin module into existing full-precision GNNs. Experiments across various domains demonstrate that the proposed method yields results superior to the state of the art.
****

STRIVE: Scene Text Replacement in Videos
Vijay Kumar B G, Jeyasri Subramanian, Varnith Chordia, Eugene Bart, Shaobo Fang, Kelly Guan, Raja Bala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14549-14558
We propose replacing scene text in videos using deep style transfer and learned photometric transformations. Building on recent progress on still image text replacement, we present extensions that alter text while preserving the appearance and motion characteristics of the original video. Compared to the problem of still image text replacement, our method addresses additional challenges introduced by video, namely effects induced by changing lighting, motion blur, diverse variations in camera-object pose over time, and preservation of temporal consistency. We parse the problem into three steps. First, the text in all frames is normalized to a frontal pose using a spatio-temporal transformer network. Second, the text is replaced in a single reference frame using a state-of-art still-image text replacement method. Finally, the new text is transferred from the reference to remaining frames using a novel learned image transformation network that captures lighting and blur effects in a temporally consistent manner. Results on syn

thetic and challenging real videos show realistic text transfer, competitive qua ntitative and qualitative performance, and superior inference speed relative to alternatives. We introduce new synthetic and real-world datasets with paired tex t objects. To the best of our knowledge this is the first attempt at deep video text replacement.
****

Disentangled High Quality Salient Object Detection
Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, Mofei Song; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2021, pp. 3580-3590
Aiming at discovering and locating most distinctive objects from visual scenes, salient object detection (SOD) plays an essential role in various computer visio n systems. Coming to the era of high resolution, SOD methods are facing new chal lenges. The major limitation of previous methods is that they try to identify th e salient regions and estimate the accurate objects boundaries simultaneously wi th a single regression task at low-resolution. This practice ignores the inheren t difference between the two difficult problems, resulting in poor detection qua lity. In this paper, we propose a novel deep learning framework for high-resolut ion SOD task, which disentangles the task into a low-resolution saliency classif ication network (LRSCN) and a high-resolution refinement network (HRRN). As a pi xel-wise classification task, LRSCN is designed to capture sufficient semantics at low-resolution to identify the definite salient, background and uncertain ima ge regions. HRRN is a regression task, which aims at accurately refining the sal iency value of pixels in the uncertain region to preserve a clear object boundar y at high-resolution with limited GPU memory. It is worth noting that by introdu cing uncertainty into the training process, our HRRN can well address the high-r esolution refinement task without using any high-resolution training data. Exten sive experiments on high-resolution saliency datasets as well as some widely use d saliency benchmarks show that the proposed method achieves superior performanc e compared to the state-of-the-art methods.
****

FREE: Feature Refinement for Generalized Zero-Shot Learning
Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, Ling S hao; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 122-131
Generalized zero-shot learning (GZSL) has achieved significant progress, with ma ny efforts dedicated to overcoming the problems of visual-semantic domain gaps a nd seen-unseen bias. However, most existing methods directly use feature extract ion models trained on ImageNet alone, ignoring the cross-dataset bias between Im ageNet and GZSL benchmarks. Such a bias inevitably results in poor-quality visua l features for GZSL tasks, which potentially limits the recognition performance on both seen and unseen classes. In this paper, we propose a simple yet effectiv e GZSL method, termed feature refinement for generalized zero-shot learning (FRE E), to tackle the above problem. FREE employs a feature refinement (FR) module t hat incorporates semantic-visual mapping into a unified generative model to refi ne the visual features of seen and unseen class samples. Furthermore, we propose a self-adaptive margin center loss (SAMC-loss) that cooperates with a semantic cycle-consistency loss to guide FR to learn class- and semantically-relevant rep resentations, and concatenate the features in FR to extract the fully refined fe atures. Extensive experiments on five benchmark datasets demonstrate the signifi cant performance gain of FREE over current state-of-the-art methods and its base line. The code is available at https://github.com/shiming-chen/FREE.
****

Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understan ding
Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautist a, Nathan Paczan, Russ Webb, Joshua M. Susskind; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2021, pp. 10912-10922
For many fundamental scene understanding tasks, it is difficult or impossible to obtain per-pixel ground truth labels from real images. We address this challeng e by introducing Hypersim, a photorealistic synthetic dataset for holistic indoo

r scene understanding. To create our dataset, we leverage a large repository of synthetic scenes created by professional artists, and we generate 77,400 images of 461 indoor scenes with detailed per-pixel labels and corresponding ground tru th geometry. Our dataset: (1) relies exclusively on publicly available 3D assets ; (2) includes complete scene geometry, material information, and lighting infor mation for every scene; (3) includes dense per-pixel semantic instance segmentat ions and complete camera information for every image; and (4) factors every imag e into diffuse reflectance, diffuse illumination, and a non-diffuse residual ter m that captures view-dependent lighting effects. We analyze our dataset at the l evel of scenes, objects, and pixels, and we analyze costs in terms of money, com putation time, and annotation effort. Remarkably, we find that it is possible to generate our entire dataset from scratch, for roughly half the cost of training a popular open-source natural language processing model. We also evaluate sim-t o-real transfer performance on two real-world scene understanding tasks - semant ic segmentation and 3D shape prediction - where we find that pre-training on our dataset significantly improves performance on both tasks, and achieves state-of -the-art performance on the most challenging Pix3D test set. All of our rendered image data, as well as all the code we used to generate our dataset and perform our experiments, is available online.

****

Self-Supervised Object Detection via Generative Image Synthesis

Siva Karthik Mustikovela, Shalini De Mello, Aayush Prakash, Umar Iqbal, Sifei Li u, Thu Nguyen-Phuoc, Carsten Rother, Jan Kautz; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2021, pp. 8609-8618

We present SSOD -- the first end-to-end analysis-by-synthesis framework with con trollable GANs for the task of self-supervised object detection. We use collecti ons of real-world images without bounding box annotations to learn to synthesize and detect objects. We leverage controllable GANs to synthesize images with pre -defined object properties and use them to train object detectors. We propose a tight end-to-end coupling of the synthesis and detection networks to optimally t rain our system. Finally, we also propose a method to optimally adapt SSOD to an intended target data without requiring labels for it. For the task of car detec tion, on the challenging KITTI and Cityscapes datasets, we show that SSOD outper forms the prior state-of-the-art purely image-based self-supervised object detec tion method Wetectron. Even without requiring any 3DCAD assets, it also surpasse s the state-of-the-art rendering-based method Meta-Sim2. Our work advances the f ield of self-supervised object detection by introducing a successful new paradig m of using controllable GAN-based image synthesis for it and by significantly im proving the base-line accuracy of the task. We open-source our code athttps://gi thub.com/NVlabs/SSOD.

****

Action-Conditioned 3D Human Motion Synthesis With Transformer VAE

Mathis Petrovich, Michael J. Black, Gül Varol; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 10985-10995

We tackle the problem of action-conditioned generation of realistic and diverse human motion sequences. In contrast to methods that complete, or extend, motion sequences, this task does not require an initial pose or sequence. Here we learn an action-aware latent representation for human motions by training a generativ e variational autoencoder (VAE). By sampling from this latent space and querying a certain duration through a series of positional encodings, we synthesize vari able-length motion sequences conditioned on a categorical action. Specifically, we design a Transformer-based architecture, ACTOR, for encoding and decoding a s equence of parametric SMPL human body models estimated from action recognition d atasets. We evaluate our approach on the NTU RGB+D, HumanAct12 and UESTC dataset s and show improvements over the state of the art. Furthermore, we present two u se cases: improving action recognition through adding our synthesized data to tr aining, and motion denoising. Code and models are available on our project page.

****

Why Approximate Matrix Square Root Outperforms Accurate SVD in Global Covariance Pooling?

Yue Song, Nicu Sebe, Wei Wang; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 1115-1123
Global Covariance Pooling (GCP) aims at exploiting the second-order statistics o f the convolutional feature. Its effectiveness has been demonstrated in boosting the classification performance of Convolutional Neural Networks (CNNs). Singula r Value Decomposition (SVD) is used in GCP to compute the matrix square root. Ho wever, the approximate matrix square root calculated using Newton-Schulz iterati on [??] outperforms the accurate one computed via SVD [??]. We empirically analy ze the reason behind the performance gap from the perspectives of data precision and gradient smoothness. Various remedies for computing smooth SVD gradients ar e investigated. Based on our observation and analyses, a hybrid training protoco l is proposed for SVD-based GCP meta-layers such that competitive performances c an be achieved against Newton-Schulz iteration. Moreover, we propose a new GCP m eta-layer that uses SVD in the forward pass, and Pade approximants in the backwa rd propagation to compute the gradients. The proposed meta-layer has been integr ated into different CNN models and achieves state-of-the-art performances on bot h large-scale and fine-grained datasets.
****

SUNet: Symmetric Undistortion Network for Rolling Shutter Correction
Bin Fan, Yuchao Dai, Mingyi He; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 4541-4550
The vast majority of modern consumer-grade cameras employ a rolling shutter mech anism, leading to image distortions if the camera moves during image acquisition . In this paper, we present a novel deep network to solve the generic rolling sh utter correction problem with two consecutive frames. Our pipeline is symmetrica lly designed to predict the global shutter image corresponding to the intermedia te time of these two frames, which is difficult for existing methods because it corresponds to a camera pose that differs most from the two frames. First, two t ime-symmetric dense undistortion flows are estimated by using well-established p rinciples: pyramidal construction, warping, and cost volume processing. Then, bo th rolling shutter images are warped into a common global shutter one in the fea ture space, respectively. Finally, a symmetric consistency constraint is constru cted in the image decoder to effectively aggregate the contextual cues of two ro lling shutter images, thereby recovering the high-quality global shutter image. Extensive experiments with both synthetic and real data from public benchmarks d emonstrate the superiority of our proposed approach over the state-of-the-art me thods.
****

DWKS: A Local Descriptor of Deformations Between Meshes and Point Clouds
Robin Magnet, Maks Ovsjanikov; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 3793-3802
We propose a novel pointwise descriptor, called DWKS, aimed at finding correspon dences across two deformable shape collections. Unlike the majority of existing descriptors, rather than capturing local geometry, DWKS captures the deformation around a point within a collection in a multi-scale and informative manner. Thi s, in turn, allows to compute inter-collection correspondences without using lan dmarks. To this end, we build upon the successful spectral WKS descriptors, but rather than using the Laplace-Beltrami operator, show that a similar constructio n can be performed on shape difference operators, that capture differences or di stortion within a collection. By leveraging the collection information our descr iptor facilitates difficult non-rigid shape matching tasks, even in the presence of strong partiality and significant deformations. We demonstrate the utility o f our approach across a range of challenging matching problems on both meshes an d point clouds. The code for this paper can be found at https://github.com/Robin Magnet/DWKS.
****

PixelPyramids: Exact Inference Models From Lossless Image Pyramids
Shweta Mahajan, Stefan Roth; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2021, pp. 6639-6648
Autoregressive models are a class of exact inference approaches with highly flex

ible functional forms, yielding state-of-the-art density estimates for natural i
mages. Yet, the sequential ordering on the dimensions makes these models computa
tionally expensive and limits their applicability to low-resolution imagery. In
this work, we propose Pixel-Pyramids, a block-autoregressive approach employing
a lossless pyramid decomposition with scale-specific representations to encode t
he joint distribution of image pixels. Crucially, it affords a sparser dependenc
y structure compared to fully autoregressive approaches. Our PixelPyramids yield
 state-of-the-art results for density estimation on various image datasets, espe
cially for high-resolution data. For CelebA-HQ 1024 x 1024, we observe that the
density estimates (in terms of bits/dim) are improved to 44% of the baseline de
spite sampling speeds superior even to easily parallelizable flow-based models.
****

Deep Blind Video Super-Resolution

Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, Jinhui Tang; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 481
1-4820

Existing video super-resolution (SR) algorithms usually assume that the blur ker
nels in the degradation process are known and do not model the blur kernels in t
he restoration. However, this assumption does not hold for blind video SR and us
ually leads to over-smoothed super-resolved frames. In this paper, we propose an
 effective blind video SR algorithm based on deep convolutional neural networks
(CNNs). Our algorithm first estimates blur kernels from low-resolution (LR) inpu
t videos. Then, with the estimated blur kernels, we develop an effective image d
econvolution method based on the image formation model of blind video SR to gene
rate intermediate latent frames so that sharp image contents can be restored wel
l. To effectively explore the information from adjacent frames, we estimate the
motion fields from LR input videos, extract features from LR videos by a feature
 extraction network, and warp the extracted features from LR inputs based on the
 motion fields. Moreover, we develop an effective sharp feature exploration meth
od which first extracts sharp features from restored intermediate latent frames
and then uses a transformation operation based on the extracted sharp features a
nd warped features from LR inputs to generate better features for HR video resto
ration. We formulate the proposed algorithm into an end-to-end trainable framewo
rk and show that it performs favorably against state-of-the-art methods.
****

Deep Relational Metric Learning

Wenzhao Zheng, Borui Zhang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2021, pp. 12065-12074

This paper presents a deep relational metric learning (DRML) framework for image
 clustering and retrieval. Most existing deep metric learning methods learn an e
mbedding space with a general objective of increasing interclass distances and d
ecreasing intraclass distances. However, the conventional losses of metric learn
ing usually suppress intraclass variations which might be helpful to identify sa
mples of unseen classes. To address this problem, we propose to adaptively learn
 an ensemble of features that characterizes an image from different aspects to m
odel both interclass and intraclass distributions. We further employ a relationa
l module to capture the correlations among each feature in the ensemble and cons
truct a graph to represent an image. We then perform relational inference on the
 graph to integrate the ensemble and obtain a relation-aware embedding to measur
e the similarities. Extensive experiments on the widely-used CUB-200-2011, Cars1
96, and Stanford Online Products datasets demonstrate that our framework improve
s existing deep metric learning methods and achieves very competitive results.
****

A Unified Objective for Novel Class Discovery

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, Elisa
 Ricci; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 9284-9292

In this paper, we study the problem of Novel Class Discovery (NCD). NCD aims at
inferring novel object categories in an unlabeled set by leveraging from prior k
nowledge of a labeled set containing different, but related classes. Existing ap

proaches tackle this problem by considering multiple objective functions, usuall y involving specialized loss terms for the labeled and the unlabeled samples res pectively, and often requiring auxiliary regularization terms. In this paper, we depart from this traditional scheme and introduce a UNified Objective function (UNO) for discovering novel classes, with the explicit purpose of favoring syner gy between supervised and unsupervised learning. Using a multi-view self-labelin g strategy, we generate pseudo-labels that can be treated homogeneously with gro und truth labels. This leads to a single classification objective operating on b oth known and unknown classes. Despite its simplicity, UNO outperforms the state of the art by a significant margin on several benchmarks (approximately +10% on CIFAR-100 and +8% on ImageNet). Our source code will be publicly available. The project page is available at: https://ncd-uno.github.io.

****

Provably Approximated Point Cloud Registration

Ibrahim Jubran, Alaa Maalouf, Ron Kimmel, Dan Feldman; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 13269-13278

The goal of the alignment problem is to align a (given) point cloud P = \ p_1,\c dots,p_n\ to another (observed) point cloud Q = \ q_1,\cdots,q_n\ . That is, to compute a rotation matrix R \in \mathbb R ^ 3 x3 and a translation vector t \i n \mathbb R ^ 3 that minimize the sum of paired distances between every transfo rmed point Rp_i-t, to its corresponding point q_i, over every i\in \br 1,\cdots, n . A harder version is the registration problem, where the correspondence is un known, and the minimum is also over all possible correspondence functions from P to Q. Algorithms such as the Iterative Closest Point (ICP) and its variants wer e suggested for these problems, but none yield a provable non-trivial approximat ion for the global optimum. We prove that there always exists a "witness" set of 3 pairs in P xQ that, via novel alignment algorithm, defines a constant factor approximation (in the worst case) to this global optimum. We then provide algori thms that recover this witness set and yield the first provable constant factor approximation for the: (i) alignment problem in O(n) expected time, and (ii) reg istration problem in polynomial time. Such small witness sets exist for many var iants including points in d-dimensional space, outlier-resistant cost functions, and different correspondence types. Extensive experimental results on real and synthetic datasets show that, in practice, our approximation constants are close to 1 and our error is up to x10 times smaller than state-of-the-art algorithms.

****

SAT: 2D Semantics Assisted Training for 3D Visual Grounding

Zhengyuan Yang, Songyang Zhang, Liwei Wang, Jiebo Luo; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 1856-1866

3D visual grounding aims at grounding a natural language description about a 3D scene, usually represented in the form of 3D point clouds, to the targeted objec t region. Point clouds are sparse, noisy, and contain limited semantic informati on compared with 2D images. These inherent limitations make the 3D visual ground ing problem more challenging. In this study, we propose 2D Semantics Assisted Tr aining (SAT) that utilizes 2D image semantics in the training stage to ease poin t-cloud-language joint representation learning and assist 3D visual grounding. T he main idea is to learn auxiliary alignments between rich, clean 2D object repr esentations and the corresponding objects or mentioned entities in 3D scenes. SA T takes 2D object semantics, i.e., object label, image feature, and 2D geometric feature, as the extra input in training but does not require such inputs during inference. By effectively utilizing 2D semantics in training, our approach boos ts the accuracy on the Nr3D dataset from 37.7% to 49.2%, which significantly sur passes the non-SAT baseline with the identical network architecture and inferenc e input. Our approach outperforms the state of the art by large margins on multi ple 3D visual grounding datasets, i.e., +10.4% absolute accuracy on Nr3D, +9.9% on Sr3D, and +5.6% on ScanRef.

****

Sample Efficient Detection and Classification of Adversarial Attacks via Self-Su pervised Embeddings

Mazda Moayeri, Soheil Feizi; Proceedings of the IEEE/CVF International Conferenc

e on Computer Vision (ICCV), 2021, pp. 7677-7686
Adversarial robustness of deep models is pivotal in ensuring safe deployment in real world settings, but most modern defenses have narrow scope and expensive costs. In this paper, we propose a self-supervised method to detect adversarial attacks and classify them to their respective threat models, based on a linear model operating on the embeddings from a pre-trained self-supervised encoder. We use a SimCLR encoder in our experiments, since we show the SimCLR embedding distance is a good proxy for human perceptibility, enabling it to encapsulate many threat models at once. We call our method SimCat since it uses SimCLR encoder to catch and categorize various types of adversarial attacks, including L_p and non-L_p evasion attacks, as well as data poisonings. The simple nature of a linear classifier makes our method efficient in both time and sample complexity. For example, on SVHN, using only five pairs of clean and adversarial examples computed with a PGD-L_inf attack, SimCat's detection accuracy is over 85%. Moreover, on ImageNet, using only 25 examples from each threat model, SimCat can classify eight different attack types such as PGD-L_2, PGD-L_inf, CW-L_2, PPGD, LPA, StAdv, ReColor, and JPEG-L_inf, with over 40% accuracy. On STL10 data, we apply SimCat as a defense against poisoning attacks, such as BP, CP, FC, CLBD, HTBD, halving the success rate while using only twenty total poisons for training. We find that the detectors generalize well to unseen threat models. Lastly, we investigate the performance of our detection method under adaptive attacks and further boost its robustness against such attacks via adversarial training.
****

Invisible Backdoor Attack With Sample-Specific Triggers
Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, Siwei Lyu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16463-16472
Recently, backdoor attacks pose a new security threat to the training process of deep neural networks (DNNs). Attackers intend to inject hidden backdoors into DNNs, such that the attacked model performs well on benign samples, whereas its prediction will be maliciously changed if hidden backdoors are activated by the attacker-defined trigger. Existing backdoor attacks usually adopt the setting that triggers are sample-agnostic, i.e., different poisoned samples contain the same trigger, resulting in that the attacks could be easily mitigated by current backdoor defenses. In this work, we explore a novel attack paradigm, where backdoor triggers are sample-specific. In our attack, we only need to modify certain training samples with invisible perturbation, while not need to manipulate other training components (e.g., training loss, and model structure) as required in many existing attacks. Specifically, inspired by the recent advance in DNN-based image steganography, we generate sample-specific invisible additive noises as backdoor triggers by encoding an attacker-specified string into benign images through an encoder-decoder network. The mapping from the string to the target label will be generated when DNNs are trained on the poisoned dataset. Extensive experiments on benchmark datasets verify the effectiveness of our method in attacking models with or without defenses.
****

Toward a Visual Concept Vocabulary for GAN Latent Space
Sarah Schwettmann, Evan Hernandez, David Bau, Samuel Klein, Jacob Andreas, Antonio Torralba; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6804-6812
A large body of recent work has identified transformations in the latent spaces of generative adversarial networks (GANs) that consistently and interpretably transform generated images. But existing techniques for identifying these transformations rely on either a fixed vocabulary of pre-specified visual concepts, or on unsupervised disentanglement techniques whose alignment with human judgments about perceptual salience is unknown. This paper introduces a new method for building open-ended vocabularies of primitive visual concepts represented in a GAN's latent space. Our approach is built from three components: (1) automatic identification of perceptually salient directions based on their layer selectivity; (2) human annotation of these directions with free-form, compositional natural lan

guage descriptions; and (3) decomposition of these annotations into a visual con cept vocabulary, consisting of distilled directions labeled with single words. E xperiments show that concepts learned with our approach are reliable and composa ble--generalizing across classes, contexts, and observers, and enabling fine-gra ined manipulation of image style and content.
****

Weakly Supervised 3D Semantic Segmentation Using Cross-Image Consensus and Inter -Voxel Affinity Relations
Xiaoyu Zhu, Jeffrey Chen, Xiangrui Zeng, Junwei Liang, Chengqi Li, Sinuo Liu, Si ma Behpour, Min Xu; Proceedings of the IEEE/CVF International Conference on Comp uter Vision (ICCV), 2021, pp. 2834-2844
We propose a novel weakly supervised approach for 3D semantic segmentation on vo lumetric images. Unlike most existing methods that require voxel-wise densely la beled training data, our weakly-supervised CIVA-Net is the first model that only needs image-level class labels as guidance to learn accurate volumetric segment ation. Our model learns from cross-image co-occurrence for integral region gener ation, and explores inter-voxel affinity relations to predict segmentation with accurate boundaries. We empirically validate our model on both simulated and rea l cryo-ET datasets. Our experiments show that CIVA-Net achieves comparable perfo rmance to the state-of-the-art models trained with stronger supervision.
****

Bootstrap Your Own Correspondences
Mohamed El Banani, Justin Johnson; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 6433-6442
Geometric feature extraction is a crucial component of point cloud registration pipelines. Recent work has demonstrated how supervised learning can be leveraged to learn better and more compact 3D features. However, those approaches' relian ce on ground-truth annotation limits their scalability. We propose BYOC: a self-supervised approach that learns visual and geometric features from RGB-D video w ithout relying on ground-truth pose or correspondence. Our key observation is th at randomly-initialized CNNs readily provide us with good correspondences; allow ing us to bootstrap the learning of both visual and geometric features. Our appr oach combines classic ideas from point cloud registration with more recent repre sentation learning approaches. We evaluate our approach on indoor scene datasets and find that our method outperforms traditional and learned descriptors, while being competitive with current state-of-the-art supervised approaches.
****

A Multi-Mode Modulator for Multi-Domain Few-Shot Classification
Yanbin Liu, Juho Lee, Linchao Zhu, Ling Chen, Humphrey Shi, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8 453-8462
Most existing few-shot classification methods only consider generalization on on e dataset (i.e., single-domain), failing to transfer across various seen and uns een domains. In this paper, we consider the more realistic multi-domain few-shot classification problem to investigate the cross-domain generalization. Two chal lenges exist in this new setting: (1) how to efficiently generate multi-domain f eature representation, and (2) how to explore domain correlations for better cro ss-domain generalization. We propose a parameter-efficient multi-mode modulator to address both challenges. First, the modulator is designed to maintain multipl e modulation parameters (one for each domain) in a single network, thus achievin g single-network multi-domain representation. Given a particular domain, domain-aware features can be efficiently generated with the well-devised separative sel ection module and cooperative query module. Second, we further divide the modula tion parameters into the domain-specific set and the domain-cooperative set to e xplore the intra-domain information and inter-domain correlations, respectively. The intra-domain information describes each domain independently to prevent neg ative interference. The inter-domain correlations guide information sharing amon g relevant domains to enrich their own representation. Moreover, unseen domains can utilize the correlations to obtain an adaptive combination of seen domains f or extrapolation. We demonstrate that the proposed multi-mode modulator achieves

state-of-the-art results on the challenging META-DATASET benchmark, especially for unseen test domains.

****

SketchAA: Abstract Representation for Abstract Sketches

Lan Yang, Kaiyue Pang, Honggang Zhang, Yi-Zhe Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10097-10106

What makes free-hand sketches appealing for humans lies with its capability as a universal tool to depict the visual world. Such flexibility at human ease, however, introduces abstract renderings that pose unique challenges to computer vision models. In this paper, we propose a purpose-made sketch representation for human sketches. The key intuition is that such representation should be abstract at design, so to accommodate the abstract nature of sketches. This is achieved by interpreting sketch abstraction on two levels: appearance and structure. We abstract sketch structure as a pre-defined coarse-to-fine visual block hierarchy, and average visual features within each block to model appearance abstraction. We then discuss three general strategies on how to exploit feature synergy across different levels of this abstraction hierarchy. The superiority of explicitly abstracting sketch representation is empirically validated on a number of sketch analysis tasks, including sketch recognition, fine-grained sketch-based image retrieval, and generative sketch healing. Our simple design not only yields strong results on all said tasks, but also offers intuitive feature granularity control to tailor for various downstream tasks. Code will be made publicly available.

****

Detecting Human-Object Relationships in Videos

Jingwei Ji, Rishi Desai, Juan Carlos Niebles; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8106-8116

We study a crucial problem in video analysis: human-object relationship detection. The majority of previous approaches are developed only for the static image scenario, without incorporating the temporal dynamics so vital to contextualizing human-object relationships. We propose a model with Intra- and Inter-Transformers, enabling joint spatial and temporal reasoning on multiple visual concepts of objects, relationships, and human poses. We find that applying attention mechanisms among features distributed spatio-temporally greatly improves our understanding of human-object relationships. Our method is validated on two datasets, Action Genome and CAD-120-EVAR, and achieves state-of-the-art performance on both of them.

****

Adaptive Curriculum Learning

Yajing Kong, Liu Liu, Jun Wang, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5067-5076

Inspired by the human learning principle that learning easier concepts first and then gradually paying more attention to harder ones, curriculum learning uses the non-uniform sampling of mini-batches according to the order of examples' difficulty. Just as a teacher adjusts the curriculum according to the learning progress of each student, a proper curriculum should be adapted to the current state of the model. Therefore, in contrast to recent works using a fixed curriculum, we devise a new curriculum learning method, Adaptive Curriculum Learning (Adaptive CL), adapting the difficulty of examples to the current state of the model. Specifically, we make use of the loss of the current model to adjust the difficulty score while retaining previous useful learned knowledge by KL divergence. Moreover, under a non-linear model and binary classification, we theoretically prove that the expected convergence rate of curriculum learning monotonically decreases with respect to the loss of a point regarding the optimal hypothesis, and monotonically increases with respect to the loss of a point regarding the current hypothesis. The analyses indicate that Adaptive CL could improve the convergence properties during the early stages of learning. Extensive experimental results demonstrate the superiority of the proposed approach over existing competitive curriculum learning methods.

****

SurfaceNet: Adversarial SVBRDF Estimation From a Single Image

Giuseppe Vecchio, Simone Palazzo, Concetto Spampinato; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12840-12848

In this paper we present SurfaceNet, an approach for estimating spatially-varying bidirectional reflectance distribution function (SVBRDF) material properties from a single image. We pose the problem as an image translation task and propose a novel patch-based generative adversarial network (GAN) that is able to produce high-quality, high-resolution surface reflectance maps. The employment of the GAN paradigm has a twofold objective: 1) allowing the model to recover finer details than standard translation models; 2) reducing the domain shift between synthetic and real data distributions in an unsupervised way. An extensive evaluation, carried out on a public benchmark of synthetic and real images under different illumination conditions, shows that SurfaceNet largely outperforms existing SVBRDF reconstruction methods, both quantitatively and qualitatively. Furthermore, SurfaceNet exhibits a re-markable ability in generating high-quality maps from real samples without any supervision at training time.

****

FloorPlanCAD: A Large-Scale CAD Drawing Dataset for Panoptic Symbol Spotting

Zhiwen Fan, Lingjie Zhu, Honghua Li, Xiaohao Chen, Siyu Zhu, Ping Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10128-10137

Access to large and diverse computer-aided design (CAD) drawings is critical for developing symbol spotting algorithms. In this paper, we present FloorPlanCAD, a large-scale real-world CAD drawing dataset containing over 10,000 floor plans, ranging from residential to commercial buildings. CAD drawings in the dataset are all represented as vector graphics, which enable us to provide line-grained annotations of 30 object categories. Equipped by such annotations, we introduce the task of panoptic symbol spotting, which requires to spot not only instances of countable things, but also the semantic of uncountable stuff. Aiming to solve this task, we propose a novel method by combining Graph Convolutional Networks (GCNs) with Convolutional Neural Networks (CNNs), which captures both non-Euclidean and Euclidean features and can be trained end-to-end. The proposed CNN-GCN method achieved state-of-the-art (SOTA) performance on the task of semantic symbol spotting, and help us build a baseline network for the panoptic symbol spotting task. Our contributions are three-fold: 1) to the best of our knowledge, the presented CAD drawing dataset is the first of its kind; 2) the panoptic symbol spotting task considers the spotting of both thing instances and stuff semantic as one recognition problem; and 3) we presented a baseline solution to the panoptic symbol spotting task based on a novel CNN-GCN method, which achieved SOTA performance on semantic symbol spotting. We believe that these contributions will boost research in related areas. The dataset and code is publicly available at https://floorplancad.github.io/.

****

TkML-AP: Adversarial Attacks to Top-k Multi-Label Learning

Shu Hu, Lipeng Ke, Xin Wang, Siwei Lyu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7649-7657

Top-k multi-label learning, which returns the top-k predicted labels from an input, has many practical applications such as image annotation, document analysis, and web search engine. However, the vulnerabilities of such algorithms with regards to dedicated adversarial perturbation attacks have not been extensively studied previously. In this work, we develop methods to create adversarial perturbations that can be used to attack top-k multi-label learning-based image annotation systems (T_kML-AP). Our methods explicitly consider the top-k ranking relation and are based on novel loss functions. Experimental evaluations on large-scale benchmark datasets including PASCAL VOC and MS COCO demonstrate the effectiveness of our methods in reducing the performance of state-of-the-art top-k multi-label learning methods, under both untargeted and targeted attacks.

****

Gradient Distribution Alignment Certificates Better Adversarial Domain Adaptation

Zhiqiang Gao, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, Chaoliang Zhong; Proceed

ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8937-8946
The latest heuristic for handling the domain shift in unsupervised domain adaptation tasks is to reduce the data distribution discrepancy using adversarial learning. Recent studies improve the conventional adversarial domain adaptation methods with discriminative information by integrating the classifier's outputs into distribution divergence measurement. However, they still suffer from the equilibrium problem of adversarial learning in which even if the discriminator is fully confused, sufficient similarity between two distributions cannot be guaranteed. To overcome this problem, we propose a novel approach named feature gradient distribution alignment (FGDA). We demonstrate the rationale of our method both theoretically and empirically. In particular, we show that the distribution discrepancy can be reduced by constraining feature gradients of two domains to have similar distributions. Meanwhile, our method enjoys a theoretical guarantee that a tighter error upper bound for target samples can be obtained than that of conventional adversarial domain adaptation methods. By integrating the proposed method with existing adversarial domain adaptation models, we achieve state-of-the-art performance on two real-world benchmark datasets.
****

Sparse-Shot Learning With Exclusive Cross-Entropy for Extremely Many Localisations
Andreas Panteli, Jonas Teuwen, Hugo Horlings, Efstratios Gavves; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2813-2823
Object localisation, in the context of regular images, often depicts objects like people or cars. In these images, there is typically a relatively small number of objects per class, which usually is manageable to annotate. However, outside the setting of regular images, we are often confronted with a different situation. In computational pathology, digitised tissue sections are extremely large images, whose dimensions quickly exceed 250'000x250'000 pixels, where relevant objects, such as tumour cells or lymphocytes can quickly number in the millions. Annotating them all is practically impossible and annotating sparsely a few, out of many more, is the only possibility. Unfortunately, learning from sparse annotations, or sparse-shot learning, clashes with standard supervised learning because what is not annotated is treated as a negative. However, assigning negative labels to what are true positives leads to confusion in the gradients and biased learning. To this end, we present exclusive cross-entropy, which slows down the biased learning by examining the second-order loss derivatives in order to drop the loss terms corresponding to likely biased terms. Experiments on nine datasets and two different localisation tasks, detection with YOLLO and segmentation with Unet, show that we obtain considerable improvements compared to cross-entropy or focal loss, while often reaching the best possible performance for the model with only 10-40% of annotations.
****

Learning Latent Architectural Distribution in Differentiable Neural Architecture Search via Variational Information Maximization
Yaoming Wang, Yuchen Liu, Wenrui Dai, Chenglin Li, Junni Zou, Hongkai Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12312-12321
Existing differentiable neural architecture search approaches simply assume the architectural distribution on each edge is independent of each other, which conflicts with the intrinsic properties of architecture. In this paper, we view the architectural distribution as the latent representation of specific data points. Then we propose Variational Information Maximization Neural Architecture Search (VIM-NAS) to leverage a simple but effective convolutional neural network to model the latent representation, and optimize for a tractable variational lower bound to the mutual information between the data points and the latent representations. VIM-NAS automatically learns a near one-hot distribution from a continuous distribution with extremely fast convergence speed, e.g., converging with one epoch. Experimental results demonstrate VIM-NAS achieves state-of-the-art perfor

mance on various search spaces, including DARTS search space, NAS-Bench-1shot1, NAS-Bench-201, and simplified search spaces S1-S4. Specifically, VIM-NAS achieves a top-1 error rate of 2.45% and 15.80% within 10 minutes on CIFAR-10 and CIFAR-100, respectively, and a top-1 error rate of 24.0% when transferred to ImageNet.

****

Motion Deblurring With Real Events

Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, Jianzhuang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2583-2592

In this paper, we propose an end-to-end learning framework for event-based motion deblurring in a self-supervised manner, where real-world events are exploited to alleviate the performance degradation caused by data inconsistency. To achieve this end, optical flows are predicted from events, with which the blurry consistency and photometric consistency are exploited to enable self-supervision on the deblurring network with real-world data. Furthermore, a piece-wise linear motion model is proposed to take into account motion non-linearities and thus leads to an accurate model for the physical formation of motion blurs in the real-world scenario. Extensive evaluation on both synthetic and real motion blur datasets demonstrates that the proposed algorithm bridges the gap between simulated and real-world motion blurs and shows remarkable performance for event-based motion deblurring in real-world scenarios.

****

Episodic Transformer for Vision-and-Language Navigation

Alexander Pashevich, Cordelia Schmid, Chen Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15942-15952

Interaction and navigation defined by natural language instructions in dynamic environments pose significant challenges for neural agents. This paper focuses on addressing two challenges: handling long sequence of subtasks, and understanding complex human instructions. We propose Episodic Transformer (E.T.), a multimodal transformer that encodes language inputs and the full episode history of visual observations and actions. To improve training, we leverage synthetic instructions as an intermediate representation that decouples understanding the visual appearance of an environment from the variations of natural language instructions. We demonstrate that encoding the history with a transformer is critical to solve compositional tasks, and that pretraining and joint training with synthetic instructions further improve the performance. Our approach sets a new state of the art on the challenging ALFRED benchmark, achieving 38.4% and 8.5% task success rates on seen and unseen test splits.

****

Change Is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery

Zhuo Zheng, Ailong Ma, Liangpei Zhang, Yanfei Zhong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15193-15202

For high spatial resolution (HSR) remote sensing images, bitemporal supervised learning always dominates change detection using many pairwise labeled bitemporal images. However, it is very expensive and time-consuming to pairwise label large-scale bitemporal HSR remote sensing images. In this paper, we propose single-temporal supervised learning (STAR) for change detection from a new perspective of exploiting object changes in unpaired images as supervisory signals. STAR enables us to train a high-accuracy change detector only using unpaired labeled images and generalize to real-world bitemporal images. To evaluate the effectiveness of STAR, we design a simple yet effective change detector called ChangeStar, which can reuse any deep semantic segmentation architecture by the ChangeMixin module. The comprehensive experimental results show that ChangeStar outperforms the baseline with a large margin under single-temporal supervision and achieves superior performance under bitemporal supervision. Code is available at https://github.com/Z-Zheng/ChangeStar.

****

Visual Saliency Transformer

Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, Junwei Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4722-4732
Existing state-of-the-art saliency detection methods heavily rely on CNN-based architectures. Alternatively, we rethink this task from a convolution-free sequence-to-sequence perspective and predict saliency by modeling long-range dependencies, which can not be achieved by convolution. Specifically, we develop a novel unified model based on a pure transformer, namely, Visual Saliency Transformer (VST), for both RGB and RGB-D salient object detection (SOD). It takes image patches as inputs and leverages the transformer to propagate global contexts among image patches. Unlike conventional architectures used in Vision Transformer (ViT), we leverage multi-level token fusion and propose a new token upsampling method under the transformer framework to get high-resolution detection results. We also develop a token-based multi-task decoder to simultaneously perform saliency and boundary detection by introducing task-related tokens and a novel patch-task-attention mechanism. Experimental results show that our model outperforms existing methods on both RGB and RGB-D SOD benchmark datasets. Most importantly, our whole framework not only provides a new perspective for the SOD field but also shows a new paradigm for transformer-based dense prediction models. Code is available at https://github.com/nnizhang/VST.
****

Event-Based Video Reconstruction Using Transformer
Wenming Weng, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2563-2572
Event cameras, which output events by detecting spatio-temporal brightness changes, bring a novel paradigm to image sensors with high dynamic range and low latency. Previous works have achieved impressive performances on event-based video reconstruction by introducing convolutional neural networks (CNNs). However, intrinsic locality of convolutional operations is not capable of modeling long-range dependency, which is crucial to many vision tasks. In this paper, we present a hybrid CNN-Transformer network for event-based video reconstruction (ET-Net), which merits the fine local information from CNN and global contexts from Transformer. In addition, we further propose a Token Pyramid Aggregation strategy to implement multi-scale token integration for relating internal and intersected semantic concepts in the token-space. Experimental results demonstrate that our proposed method achieves superior performance over state-of-the-art methods on multiple real-world event datasets. The code is available at https://github.com/Warranweng/ET-Net
****

Naturalistic Physical Adversarial Patch for Object Detectors
Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, Wen-Huang Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7848-7857
Most prior works on physical adversarial attacks mainly focus on the attack performance but seldom enforce any restrictions over the appearance of the generated adversarial patches. This leads to conspicuous and attention-grabbing patterns for the generated patches which can be easily identified by humans. To address this issue, we propose a method to craft physical adversarial patches for object detectors by leveraging the learned image manifold of a pretrained generative adversarial network (GAN) (e.g., BigGAN and StyleGAN) upon real-world images. Through sampling the optimal image from the GAN, our method can generate natural looking adversarial patches while maintaining high attack performance. With extensive experiments on both digital and physical domains and several independent subjective surveys, the results show that our proposed method produces significantly more realistic and natural looking patches than several state-of-the-art baselines while achieving competitive attack performance.
****

Attentive and Contrastive Learning for Joint Depth and Motion Field Estimation
Seokju Lee, Francois Rameau, Fei Pan, In So Kweon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4862-4871
Estimating the motion of the camera together with the 3D structure of the scene

from a monocular vision system is a complex task that often relies on the so-called scene rigidity assumption. When observing a dynamic environment, this assumption is violated which leads to an ambiguity between the ego-motion of the camera and the motion of the objects. To solve this problem, we present a self-supervised learning framework for 3D object motion field estimation from monocular videos. Our contributions are two-fold. First, we propose a two-stage projection pipeline to explicitly disentangle the camera ego-motion and the object motions with dynamics attention module, called DAM. Specifically, we design an integrated motion model that estimates the motion of the camera and object in the first and second warping stages, respectively, controlled by the attention module through a shared motion encoder. Second, we propose an object motion field estimation through contrastive sample consensus, called CSAC, taking advantage of weak semantic prior (bounding box from an object detector) and geometric constraints (each object respects the rigid body motion model). Experiments on KITTI, Cityscapes, and Waymo Open Dataset demonstrate the relevance of our approach and show that our method outperforms state-of-the-art algorithms for the tasks of self-supervised monocular depth estimation, object motion segmentation, monocular scene flow estimation, and visual odometry.

****

Graspness Discovery in Clutters for Fast and Accurate Grasp Detection

Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15964-15973

Efficient and robust grasp pose detection is vital for robotic manipulation. For general 6 DoF grasping, conventional methods treat all points in a scene equally and usually adopt uniform sampling to select grasp candidates. However, we discover that ignoring where to grasp greatly harms the speed and accuracy of current grasp pose detection methods. In this paper, we propose "graspness", a quality based on geometry cues that distinguishes graspable area in cluttered scenes. A look-ahead searching method is proposed for measuring the graspness and statistical results justify the rationality of our method. To quickly detect graspness in practice, we develop a neural network named graspness model to approximate the searching process. Extensive experiments verify the stability, generality and effectiveness of our graspness model, allowing it to be used as a plug-and-play module for different methods. A large improvement in accuracy is witnessed for various previous methods after equipping our graspness model. Moreover, we develop GSNet, an end-to-end network that incorporates our graspness model for early filtering of low-quality predictions. Experiments on a large-scale benchmark, GraspNet-1Billion, show that our method outperforms previous arts by a large margin (30+ AP) and achieves a high inference speed. Our code and model will be made publicly available.

****

Gradient Normalization for Generative Adversarial Networks

Yi-Lun Wu, Hong-Han Shuai, Zhi-Rui Tam, Hong-Yu Chiu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6373-6382

In this paper, we propose a novel normalization method called gradient normalization (GN) to tackle the training instability of Generative Adversarial Networks (GANs) caused by the sharp gradient space. Unlike existing work such as gradient penalty and spectral normalization, the proposed GN only imposes a hard 1-Lipschitz constraint on the discriminator function, which increases the capacity of the discriminator. Moreover, the proposed gradient normalization can be applied to different GAN architectures with little modification. Extensive experiments on four datasets show that GANs trained with gradient normalization outperform existing methods in terms of both Frechet Inception Distance and Inception Score.

****

Clustering by Maximizing Mutual Information Across Views

Kien Do, Truyen Tran, Svetha Venkatesh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9928-9938

We propose a novel framework for image clustering that incorporates joint representation learning and clustering. Our method consists of two heads that share th

e same backbone network - a "representation learning" head and a "clustering" he ad. The "representation learning" head captures fine-grained patterns of objects at the instance level which serve as clues for the "clustering" head to extract coarse-grain information that separates objects into clusters. The whole model is trained in an end-to-end manner by minimizing the weighted sum of two sample-oriented contrastive losses applied to the outputs of the two heads. To ensure t hat the contrastive loss corresponding to the "clustering" head is optimal, we i ntroduce a novel critic function called "log-of-dot-product". Extensive experime ntal results demonstrate that our method significantly outperforms state-of-the-art single-stage clustering methods across a variety of image datasets, improvin g over the best baseline by about 5-7% in accuracy on CIFAR10/20, STL10, and Ima geNet-Dogs. Further, the "two-stage" variant of our method also achieves better results than baselines on three challenging ImageNet subsets.

****

SIGN: Spatial-Information Incorporated Generative Network for Generalized Zero-S hot Semantic Segmentation

Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, Wael Abd-Almageed; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 955 6-9566

Unlike conventional zero-shot classification, zero-shot semantic segmentation pr edicts a class label at the pixel level instead of the image level. When solving zero-shot semantic segmentation problems, the need for pixel-level prediction w ith surrounding context motivates us to incorporate spatial information using po sitional encoding. We improve standard positional encoding by introducing the co ncept of Relative Positional Encoding, which integrates spatial information at t he feature level and can handle arbitrary image sizes. Furthermore, while self-t raining is widely used in zero-shot semantic segmentation to generate pseudo-lab els, we propose a new knowledge-distillation-inspired self-training strategy, na mely Annealed Self-Training, which can automatically assign different importance to pseudo-labels to improve performance. We systematically study the proposed R elative Positional Encoding and Annealed Self-Training in a comprehensive experi mental evaluation, and our empirical results confirm the effectiveness of our me thod on three benchmark datasets.

****

Learning With Privileged Tasks

Yuru Song, Zan Lou, Shan You, Erkun Yang, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10685-10694

Multi-objective multi-task learning aims to boost the performance of all tasks b y leveraging their correlation and conflict appropriately. Nevertheless, in real practice, users may have preference for certain tasks, and other tasks simply s erve as privileged or auxiliary tasks to assist the training of target tasks. Th e privileged tasks thus possess less or even no priority in the final task asses sment by users. Motivated by this, we propose a privileged multiple descent algo rithm to arbitrate the learning of target tasks and privileged tasks. Concretely , we introduce a privileged parameter so that the optimization direction does no t necessarily follow the gradient from the privileged tasks, but concentrates mo re on the target tasks. Besides, we also encourage a priority parameter for the target tasks to control the potential distraction of optimization direction from the privileged tasks. In this way, the optimization direction can be more aggre ssively determined by weighting the gradients among target and privileged tasks, and thus highlight more the performance of target tasks under the unified multi -task learning context. Extensive experiments on synthetic and real-world datase ts indicate that our method can achieve versatile Pareto solutions under varying preference for the target tasks.

****

Modelling Neighbor Relation in Joint Space-Time Graph for Video Correspondence L earning

Zixu Zhao, Yueming Jin, Pheng-Ann Heng; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2021, pp. 9960-9969

This paper presents a self-supervised method for learning reliable visual correspondence from unlabeled videos. We formulate the correspondence as finding paths in a joint space-time graph, where nodes are grid patches sampled from frames, and are linked by two type of edges: (i) neighbor relations that determine the aggregation strength from intra-frame neighbors in space, and (ii) similarity relations that indicate the transition probability of inter-frame paths across time. Leveraging the cycle-consistency in videos, our contrastive learning objective discriminates dynamic objects from both their neighboring views and temporal views. Compared with prior works, our approach actively explores the neighbor relations of central instances to learn a latent association between center-neighbor pairs (eg, "hand -- arm") across time, thus improving the instance discrimination. Without fine-tuning, our learned representation outperforms the state-of-the-art self-supervised methods on a variety of visual tasks including video object propagation, part propagation, and pose keypoint tracking. Our self-supervised method also surpasses some fully supervised algorithms designed for the specific tasks.

****

DiagViB-6: A Diagnostic Benchmark Suite for Vision Models in the Presence of Shortcut and Generalization Opportunities

Elias Eulig, Piyapat Saranrittichai, Chaithanya Kumar Mummadi, Kilian Rambach, William Beluch, Xiahan Shi, Volker Fischer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10655-10664

Common deep neural networks (DNNs) for image classification have been shown to rely on shortcut opportunities (SO) in the form of predictive and easy-to-represent visual factors. This is known as shortcut learning and leads to impaired generalization. In this work, we show that common DNNs also suffer from shortcut learning when predicting only basic visual object factors of variation (FoV) such as shape, color, or texture. We argue that besides shortcut opportunities, generalization opportunities (GO) are also an inherent part of real-world vision data and arise from partial independence between predicted classes and FoVs. We also argue that it is necessary for DNNs to exploit GO to overcome shortcut learning. Our core contribution is to introduce the Diagnostic Vision Benchmark suite DiagViB-6, which includes datasets and metrics to study a network's shortcut vulnerability and generalization capability for six independent FoV. In particular, DiagViB-6 allows controlling the type and degree of SO and GO in a dataset. We benchmark a wide range of popular vision architectures and show that they can exploit GO only to a limited extent.

****

ASMR: Learning Attribute-Based Person Search With Adaptive Semantic Margin Regularizer

Boseung Jeong, Jicheol Park, Suha Kwak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12016-12025

Attribute-based person search is the task of finding person images that are best matched with a set of text attributes given as query. The main challenge of this task is the large modality gap between attributes and images. To reduce the gap, we present a new loss for learning cross-modal embeddings in the context of attribute-based person search. We regard a set of attributes as a category of people sharing the same traits. In a joint embedding space of the two modalities, our loss pulls images close to their person categories for modality alignment. More importantly, it pushes apart a pair of person categories by a margin determined adaptively by their semantic distance, where the distance metric is learned end-to-end so that the loss considers importance of each attribute when relating person categories. Our loss guided by the adaptive semantic margin leads to more discriminative and semantically well-arranged distributions of person images. As a consequence, it enables a simple embedding model to achieve state-of-the-art records on public benchmarks without bells and whistles.

****

Learning Inner-Group Relations on Point Clouds

Haoxi Ran, Wei Zhuo, Jun Liu, Li Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15477-15487

The prevalence of relation networks in computer vision is in stark contrast to u
nderexplored point-based methods. In this paper, we explore the possibilities of
 local relation operators and survey their feasibility. We propose a scalable an
d efficient module, called group relation aggregator. The module computes a feat
ure of a group based on the aggregation of the features of the inner-group point
s weighted by geometric relations and semantic relations. For convenience, we ge
neralize groupwise operations to assemble this module. We adopt this module to d
esign our RPNet. We further verify the expandability of RPNet, in terms of both
depth and width, on the tasks of classification and segmentation. Surprisingly,
empirical results show that wider RPNet fits for classification, while deeper RP
Net works better on segmentation. RPNet achieves state-of-the-art for classifica
tion and segmentation on challenging benchmarks. We also compare our local aggre
gator with PointNet++, with around 30% parameters and 50% computation saving. Fi
nally, we conduct experiments to reveal the robustness of RPNet with regard to r
igid transformation and noises.
****

Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows
Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baini
ng Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 10012-10022
This paper presents a new vision Transformer, called Swin Transformer, that capa
bly serves as a general-purpose backbone for computer vision. Challenges in adap
ting Transformer from language to vision arise from differences between the two
domains, such as large variations in the scale of visual entities and the high r
esolution of pixels in images compared to words in text. To address these differ
ences, we propose a hierarchical Transformer whose representation is computed wi
th Shifted windows. The shifted windowing scheme brings greater efficiency by li
miting self-attention computation to non-overlapping local windows while also al
lowing for cross-window connection. This hierarchical architecture has the flexi
bility to model at various scales and has linear computational complexity with r
espect to image size. These qualities of Swin Transformer make it compatible wit
h a broad range of vision tasks, including image classification (87.3 top-1 accu
racy on ImageNet-1K) and dense prediction tasks such as object detection (58.7 b
ox AP and 51.1 mask AP on COCO test-dev) and semantic segmentation (53.5 mIoU on
 ADE20K val). Its performance surpasses the previous state-of-the-art by a large
 margin of +2.7 box AP and +2.6 mask AP on COCO, and +3.2 mIoU on ADE20K, demons
trating the potential of Transformer-based models as vision backbones. The hiera
rchical design and the shifted window approach also prove beneficial for all-MLP
 architectures. The code and models are publicly available at https://github.com
/microsoft/Swin-Transformer.
****

Dynamic Cross Feature Fusion for Remote Sensing Pansharpening
Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Tian-Jing Zhang; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14687-14696
Deep Convolution Neural Networks have been adopted for pansharpening and achieve
d state-of-the-art performance. However, most of the existing works mainly focus
 on single-scale feature fusion, which leads to failure in fully considering rel
ationships of information between high-level semantics and low-level features, d
espite the network is deep enough. In this paper, we propose a dynamic cross fea
ture fusion network (DCFNet) for pansharpening. Specifically, DCFNet contains mu
ltiple parallel branches, including a high-resolution branch served as the backb
one, and the low-resolution branches progressively supplemented into the backbon
e. Thus our DCFNet can represent the overall information well. In order to enhan
ce the relationships of inter-branches, dynamic cross feature transfers are embe
dded into multiple branches to obtain high-resolution representations. Then cont
extualized features will be learned to improve the fusion of information. Experi
mental results indicate that DCFNet significantly outperforms the prior arts in
both quantitative indicators and visual qualities.
****

Free-Form Description Guided 3D Visual Graph Network for Object Grounding in Poi

nt Cloud
Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Z hang, Yaonan Wang, Ajmal Mian; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2021, pp. 3722-3731
3D object grounding aims to locate the most relevant target object in a raw poin t cloud scene based on a free-form language description. Understanding complex a nd diverse descriptions, and lifting them directly to a point cloud is a new and challenging topic due to the irregular and sparse nature of point clouds. There are three main challenges in 3D object grounding: to find the main focus in the complex and diverse description; to understand the point cloud scene; and to lo cate the target object. In this paper, we address all three challenges. Firstly, we propose a language scene graph module to capture the rich structure and long -distance phrase correlations. Secondly, we introduce a multi-level 3D proposal relation graph module to extract the object-object and object-scene co-occurrenc e relationships, and strengthen the visual features of the initial proposals. La stly, we develop a description guided 3D visual graph module to encode global co ntexts of phrases and proposals by a nodes matching strategy. Extensive experime nts on challenging benchmark datasets (ScanRefer and Nr3D) show that our algorit hm outperforms existing state-of-the-art. Our code is available at https://githu b.com/PNXD/FFL-3DOG.
****

The Devil Is in the Task: Exploiting Reciprocal Appearance-Localization Features for Monocular 3D Object Detection
Zhikang Zou, Xiaoqing Ye, Liang Du, Xianhui Cheng, Xiao Tan, Li Zhang, Jianfeng Feng, Xiangyang Xue, Errui Ding; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 2713-2722
Low-cost monocular 3D object detection plays a fundamental role in autonomous dr iving, whereas its accuracy is still far from satisfactory. Our objective is to dig into the 3D object detection task and reformulate it as the sub-tasks of obj ect localization and appearance perception, which benefits to a deep excavation of reciprocal information underlying the entire task. We introduce a Dynamic Fea ture Reflecting Network, named DFR-Net, which contains two novel standalone modu les: (i) the Appearance-Localization Feature Reflecting module (ALFR) that first separates task-specific features and then self-mutually reflects the reciprocal features; (ii) the Dynamic Intra-Trading module (DIT) that adaptively realigns the training processes of various sub-tasks via a self-learning manner. Extensiv e experiments on the challenging KITTI dataset demonstrate the effectiveness and generalization of DFR-Net. We rank 1st among all the monocular 3D object detect ors in the KITTI test set (till March 16th, 2021). The proposed method is also e asy to be plug-and-play in many cutting-edge 3D detection frameworks at negligib le cost to boost performance. The code will be made publicly available.
****

SimROD: A Simple Adaptation Method for Robust Object Detection
Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, Yong Zh ang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 3570-3579
This paper presents a Simple and effective unsupervised adaptation method for Ro bust Object Detection (SimROD). To overcome the challenging issues of domain shi ft and pseudo-label noise, our method integrates a novel domain-centric data aug mentation, a gradual self-labeling adaptation procedure, and a teacher-guided fi ne-tuning mechanism. Using our method, target domain samples can be leveraged to adapt object detection models without changing the model architecture or genera ting synthetic data. When applied to image corruptions and high-level cross-doma in adaptation benchmarks, our method outperforms prior baselines on multiple dom ain adaptation benchmarks. SimROD achieves new state-of-the-art on standard real -to-synthetic and cross-camera setup benchmarks. On the image corruption benchma rk, models adapted with our method achieved a relative robustness improvement of 15-25% AP50 on Pascal-C and 5-6% AP on COCO-C and Cityscapes-C. On the cross-do main benchmark, our method outperformed the best baseline performance by up to 8 % and 4% AP50 on Comic and Watercolor respectively.

****
Gated3D: Monocular 3D Object Detection From Temporal Illumination Cues
Frank Julca-Aguilar, Jason Taylor, Mario Bijelic, Fahim Mannan, Ethan Tseng, Felix Heide; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2938-2948
Today's state-of-the-art methods for 3D object detection are based on lidar, stereo, or monocular cameras. Lidar-based methods achieve the best accuracy, but have a large footprint, high cost, and mechanically-limited angular sampling rates, resulting in low spatial resolution at long ranges. Recent approaches using low-cost monocular or stereo cameras promise to overcome these limitations but struggle in low-light or low-contrast regions as they rely on passive CMOS sensors. We propose a novel 3D object detection modality that exploits temporal illumination cues from a low-cost monocular gated imager. We introduce a novel deep detection architecture, Gated3D, that is tailored to temporal illumination cues in gated images. This modality allows us to exploit mature 2D object feature extractors that guide the 3D predictions through a frustum segment estimation. We assess the proposed method experimentally on a 3D detection dataset that includes gated images captured over 10,000 km of driving data. We validate that our method outperforms state-of-the-art monocular and stereo methods, opening up a new sensor modality as an avenue to replace lidar in autonomous driving. https://light.princeton.edu/gated3d
****
iMAP: Implicit Mapping and Positioning in Real-Time
Edgar Sucar, Shikun Liu, Joseph Ortiz, Andrew J. Davison; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6229-6238
We show for the first time that a multilayer perceptron (MLP) can serve as the only scene representation in a real-time SLAM system for a handheld RGB-D camera. Our network is trained in live operation without prior data, building a dense, scene-specific implicit 3D model of occupancy and colour which is also immediately used for tracking. Achieving real-time SLAM via continual training of a neural network against a live image stream requires significant innovation. Our iMAP algorithm uses a keyframe structure and multi-processing computation flow, with dynamic information-guided pixel sampling for speed, with tracking at 10 Hz and global map updating at 2 Hz. The advantages of an implicit MLP over standard dense SLAM techniques include efficient geometry representation with automatic detail control and smooth, plausible filling-in of unobserved regions such as the back surfaces of objects.
****
Conditional Diffusion for Interactive Segmentation
Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, Manni Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7345-7354
In click-based interactive segmentation, the mask extraction process is dictated by positive/negative user clicks; however, most existing methods do not fully exploit the user cues, requiring excessive numbers of clicks for satisfactory results. We propose Conditional Diffusion Network(CDNet), which propagates labeled representations from clicks to conditioned destinations with two levels of affinities: Feature Diffusion Module (FDM) spreads features from clicks to potential target regions with global similarity; Pixel Diffusion Module (PDM) diffuses the predicted logits of clicks within locally connected regions. Thus, the information inferred by user clicks could be generalized to proper destinations. In addition, we put forward Diversified Training(DT), which reduces the optimization ambiguity caused by click simulation. With FDM,PDM and DT, CDNet could better understand user's intentions and make better predictions with limited interactions. CDNet achieves state-of-the-art performance on several benchmarks.
****
Semi-Supervised Active Learning for Semi-Supervised Models: Exploit Adversarial Examples With Graph-Based Virtual Labels
Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, Yueting Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2896-2905

The performance of computer vision models significantly improves with more label ed data. However, the acquisition of labeled data is limited by the high cost. T o mitigate the reliance on large labeled datasets, active learning (AL) and semi -supervised learning (SSL) are frequently adopted. Although current mainstream m ethods begin to combine SSL and AL (SSL-AL) to excavate the diverse expressions of unlabeled samples, these methods' fully supervised task models are still trai ned only with labeled data. Besides, these method's SSL-AL frameworks suffer fro m mismatch problems. Here, we propose a graph-based SSL-AL framework to unleash the SSL task models' power and make an effective SSL-AL interaction. In the fram ework, SSL leverages graph-based label propagation to deliver virtual labels to unlabeled samples, rendering AL samples' structural distribution and boosting AL . AL finds samples near the clusters' boundary to help SSL perform better label propagation by exploiting adversarial examples. The information exchange in the closed-loop realizes mutual enhancement of SSL and AL. Experimental results show that our method outperforms the state-of-the-art methods against classification and segmentation benchmarks.

****

RobustNav: Towards Benchmarking Robustness in Embodied Navigation

Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, Aniruddha Kembhavi; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 15691-15700

As an attempt towards assessing the robustness of embodied navigation agents, we propose RobustNav, a framework to quantify the performance of embodied navigati on agents when exposed to a wide variety of visual-- affecting RGB inputs -- and dynamics -- affecting transition dynamics -- corruptions. Most recent efforts i n visual navigation have typically focused on generalizing to novel target envir onments with similar appearance and dynamics characteristics. With RobustNav, we find that some standard embodied navigation agents significantly underperform (or fail) in the presence of visual or dynamics corruptions. We systematically an alyze the kind of idiosyncrasies that emerge in the behavior of such agents when operating under corruptions. Finally, for visual corruptions in RobustNav, we s how that while standard techniques to improve robustness such as data-augmentati on and self-supervised adaptation offer some zero-shot resistance and improvemen ts in navigation performance, there is still a long way to go in terms of recove ring lost performance relative to clean "non-corrupt" settings, warranting more research in this direction. Our code is available at https://github.com/allenai/ robustnav.

****

Conditional Variational Capsule Network for Open Set Recognition

Yunrui Guo, Guglielmo Camporese, Wenjing Yang, Alessandro Sperduti, Lamberto Bal lan; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 103-111

In open set recognition, a classifier has to detect unknown classes that are not known at training time. In order to recognize new categories, the classifier ha s to project the input samples of known classes in very compact and separated re gions of the features space for discriminating samples of unknown classes. Recen tly proposed Capsule Networks have shown to outperform alternatives in many fiel ds, particularly in image recognition, however they have not been fully applied yet to open-set recognition. In capsule networks, scalar neurons are replaced by capsule vectors or matrices, whose entries represent different properties of ob jects. In our proposal, during training, capsules features of the same known cla ss are encouraged to match a pre-defined gaussian, one for each class. To this e nd, we use the variational autoencoder framework, with a set of gaussian priors as the approximation for the posterior distribution. In this way, we are able to control the compactness of the features of the same class around the center of the gaussians, thus controlling the ability of the classifier in detecting sampl es from unknown classes. We conducted several experiments and ablation of our mo del, obtaining state of the art results on different datasets in the open set re cognition and unknown detection tasks.

****

Towards Real-World Prohibited Item Detection: A Large-Scale X-Ray Benchmark
Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, Yanjun Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5412-5421
Automatic security inspection using computer vision technology is a challenging task in real-world scenarios due to various factors, including intra-class variance, class imbalance, and occlusion. Most of the previous methods rarely solve the cases that the prohibited items are deliberately hidden in messy objects due to the lack of large-scale datasets, restricted their applications in real-world scenarios. Towards real-world prohibited item detection, we collect a large-scale dataset, named as PIDray, which covers various cases in real-world scenarios for prohibited item detection, especially for deliberately hidden items. With an intensive amount of effort, our dataset contains 12 categories of prohibited items in 47,677 X-ray images with high-quality annotated segmentation masks and bounding boxes. To the best of our knowledge, it is the largest prohibited items detection dataset to date. Meanwhile, we design the selective dense attention network (SDANet) to construct a strong baseline, which consists of the dense attention module and the dependency refinement module. The dense attention module formed by the spatial and channel-wise dense attentions, is designed to learn the discriminative features to boost the performance. The dependency refinement module is used to exploit the dependencies of multi-scale features. Extensive experiments conducted on the collected PIDray dataset demonstrate that the proposed method performs favorably against the state-of-the-art methods, especially for detecting the deliberately hidden items.
****

Let's See Clearly: Contaminant Artifact Removal for Moving Cameras
Xiaoyu Li, Bo Zhang, Jing Liao, Pedro V. Sander; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2011-2020
Contaminants such as dust, dirt and moisture adhering to the camera lens can greatly affect the quality and clarity of the resulting image or video. In this paper, we propose a video restoration method to automatically remove these contaminants and produce a clean video. Our approach first seeks to detect attention maps that indicate the regions that need to be restored. In order to leverage the corresponding clean pixels from adjacent frames, we propose a flow completion module to hallucinate the flow of the background scene to the attention regions degraded by the contaminants. Guided by the attention maps and completed flows, we propose a recurrent technique to restore the input frame by fetching clean pixels from adjacent frames. Finally, a multi-frame processing stage is used to further process the entire video sequence in order to enforce temporal consistency. The entire network is trained on a synthetic dataset that approximates the physical lighting properties of contaminant artifacts. This new dataset and our novel framework lead to our method that is able to address different contaminants and outperforms competitive restoration approaches both qualitatively and quantitatively.
****

Generating Attribution Maps With Disentangled Masked Backpropagation
Adria Ruiz, Antonio Agudo, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 905-914
Attribution map visualization has arisen as one of the most effective techniques to understand the underlying inference process of Convolutional Neural Networks. In this task, the goal is to compute an score for each image pixel related to its contribution to the network output. In this paper, we introduce Disentangled Masked Backpropagation (DMBP), a novel gradient-based method that leverages on the piecewise linear nature of ReLU networks to decompose the model function into different linear mappings. This decomposition aims to disentangle the attribution maps into positive, negative and nuisance factors by learning a set of variables masking the contribution of each filter during back-propagation. A thorough evaluation over standard architectures (ResNet50 and VGG16) and benchmark datasets (PASCAL VOC and ImageNet) demonstrates that DMBP generates more visually interpretable attribution maps than previous approaches. Additionally, we quantitat

ively show that the maps produced by our method are more consistent with the tru
e contribution of each pixel to the final network output.
****

A Simple Baseline for Weakly-Supervised Scene Graph Generation
Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, Chenliang Xu; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 16393-16402
We investigate the weakly-supervised scene graph generation, which is a challeng
ing task since no correspondence of label and object is provided. The previous w
ork regards such correspondence as a latent variable which is iteratively update
d via nested optimization of the scene graph generation objective. However, we f
urther reduce the complexity by decoupling it into an efficient first-order grap
h matching module optimized via contrastive learning to obtain such corresponden
ce, which is used to train a standard scene graph generation model. The extensiv
e experiments show that such a simple pipeline can significantly surpass the pre
vious state-of-the-art by more than 30% on the Visual Genome dataset, both in te
rms of graph matching accuracy and scene graph quality. We believe this work ser
ves as a strong baseline for future research.
****

Self-Supervised Vessel Segmentation via Adversarial Learning
Yuxin Ma, Yang Hua, Hanming Deng, Tao Song, Hao Wang, Zhengui Xue, Heng Cao, Ruh
ui Ma, Haibing Guan; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2021, pp. 7536-7545
Vessel segmentation is critically essential for diagnosinga series of diseases,
e.g., coronary artery disease and retinal disease. However, annotating vessel se
gmentation maps of medical images is notoriously challenging due to the tiny and
 complex vessel structures, leading to insufficient available annotated datasets
 for existing supervised methods and domain adaptation methods. The subtle struc
tures and confusing background of medical images further suppress the efficacy o
f unsupervised methods. In this paper, we propose a self-supervised vessel segme
ntation method via adversarial learning. Our method learns vessel representation
s by training an attention-guided generator and a segmentation generator to simu
ltaneously synthesize fake vessels and segment vessels out of coronary angiogram
s. To support the research, we also build the first X-ray angiography coronary v
essel segmentation dataset, named XCAD. We evaluate our method extensively on mu
ltiple vessel segmentation datasets, including the XCAD dataset, the DRIVE datas
et,and the STARE dataset. The experimental results show our method suppresses un
supervised methods significantly and achieves competitive performance compared w
ith supervised methods and traditional methods.
****

3DStyleNet: Creating 3D Shapes With Geometric and Texture Style Variations
Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, Sanja Fidler; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1245
6-12465
We propose a method to create plausible geometric and texture style variations o
f 3D objects in the quest to democratize 3D content creation. Given a pair of te
xtured source and target objects, our method predicts a part-aware affine transf
ormation field that naturally warps the source shape to imitate the overall geom
etric style of the target. In addition, the texture style of the target is trans
ferred to the warped source object with the help of a multi-view differentiable
renderer. Our model, 3DStyleNet, is composed of two sub-networks trained in two
stages. First, the geometric style network is trained on a large set of untextur
ed 3D shapes. Second, we jointly optimize our geometric style network and a pre-
trained image style transfer network with losses defined over both the geometry
and the rendering of the result. Given a small set of high-quality textured obje
cts, our method can create many novel stylized shapes, resulting in effortless 3
D content creation and style-ware data augmentation. We showcase our approach qu
alitatively on 3D content stylization, and provide user studies to validate the
quality of our results. In addition, our method can serve as a valuable tool to
create 3D data augmentations for computer vision tasks. Extensive quantitative a
nalysis shows that 3DStyleNet outperforms alternative data augmentation techniqu

es for the downstream task of single-image 3D reconstruction.
****

NeRD: Neural Reflectance Decomposition From Image Collections
Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, Hendrik P.A. Lensch; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12684-12694
Decomposing a scene into its shape, reflectance, and illumination is a challenging but important problem in computer vision and graphics. This problem is inherently more challenging when the illumination is not a single light source under laboratory conditions but is instead an unconstrained environmental illumination. Though recent work has shown that implicit representations can be used to model the radiance field of an object, most of these techniques only enable view synthesis and not relighting. Additionally, evaluating these radiance fields is resource and time-intensive. We propose a neural reflectance decomposition (NeRD) technique that uses physically-based rendering to decompose the scene into spatially varying BRDF material properties. In contrast to existing techniques, our input images can be captured under different illumination conditions. In addition, we also propose techniques to convert the learned reflectance volume into a relightable textured mesh enabling fast real-time rendering with novel illuminations. We demonstrate the potential of the proposed approach with experiments on both synthetic and real datasets, where we are able to obtain high-quality relightable 3D assets from image collections. The datasets and code are available at the project page: https://markboss.me/publication/2021-nerd/
****

Deep Hybrid Self-Prior for Full 3D Mesh Generation
Xingkui Wei, Zhengqing Chen, Yanwei Fu, Zhaopeng Cui, Yinda Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5805-5814
We present a deep learning pipeline that leverages network self-prior to recover a full 3D model consisting of both a triangular mesh and a texture map from the colored 3D point cloud. Different from previous methods either exploiting 2D self-prior for image editing or 3D self-prior for pure surface reconstruction, we propose to exploit a novel hybrid 2D-3D self-prior in deep neural networks to significantly improve the geometry quality and produce a high-resolution texture map, which is typically missing from the output of commodity-level 3D scanners. In particular, we first generate an initial mesh using a 3D convolutional neural network with 3D self-prior, and then encode both 3D information and color information in the 2D UV atlas, which is further refined by 2D convolutional neural networks with the self-prior. In this way, both 2D and 3D self-priors are utilized for the mesh and texture recovery. Experiments show that, without the need of any additional training data, our method recovers the 3D textured mesh model of high quality from sparse input, and outperforms the state-of-the-art methods in terms of both the geometry and texture quality.
****

MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction
Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, Guiqing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11467-11476
Human motion prediction is a challenging task due to the stochasticity and aperiodicity of future poses. Recently, graph convolutional network has been proven to be very effective to learn dynamic relations among pose joints, which is helpful for pose prediction. On the other hand, one can abstract a human pose recursively to obtain a set of poses at multiple scales. With the increase of the abstraction level, the motion of the pose becomes more stable, which benefits pose prediction too. In this paper, we propose a novel Multi-Scale Residual Graph Convolution Network (MSR-GCN) for human pose prediction task in the manner of end-to-end. The GCNs are used to extract features from fine to coarse scale and then from coarse to fine scale. The extracted features at each scale are then combined and decoded to obtain the residuals between the input and target poses. Intermed

iate supervisions are imposed on all the predicted poses, which enforces the network to learn more representative features. Our proposed approach is evaluated on two standard benchmark datasets, i.e., the Human3.6M dataset and the CMU Mocap dataset. Experimental results demonstrate that our method outperforms the state-of-the-art approaches. Code and pre-trained models are available at https://github.com/Droliven/MSRGCN.
****

Super Resolve Dynamic Scene From Continuous Spike Streams
Jing Zhao, Jiyu Xie, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Tiejun Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2533-2542
Recently, a novel retina-inspired camera, namely spike camera, has shown great potential for recording high-speed dynamic scenes. Unlike the conventional digital cameras that compact the visual information within the exposure interval into a single snapshot, the spike camera continuously outputs binary spike streams to record the dynamic scenes, yielding a very high temporal resolution. Most of the existing reconstruction methods for spike camera focus on reconstructing images with the same resolution as spike camera. However, as a trade-off of high temporal resolution, the spatial resolution of spike camera is limited, resulting in inferior details of the reconstruction. To address this issue, we develop a spike camera super-resolution framework, aiming to super resolve high-resolution intensity images from the low-resolution binary spike streams. Due to the relative motion between the camera and the objects to capture, the spikes fired by the same sensor pixel no longer describes the same points in the external scene. In this paper, we properly exploit the relative motion and derive the relationship between light intensity and each spike, so as to recover the external scene with both high temporal and high spatial resolution. Experimental results demonstrate that the proposed method can reconstruct pleasant high-resolution images from low-resolution spike streams.
****

Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation
Beibei Lin, Shunli Zhang, Xin Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14648-14656
Gait recognition is one of the most important biometric technologies and has been applied in many fields. Recent gait recognition frameworks represent each gait frame by descriptors extracted from either global appearances or local regions of humans. However, the representations based on global information often neglect the details of the gait frame, while local region based descriptors cannot capture the relations among neighboring regions, thus reducing their discriminativeness. In this paper, we propose a novel feature extraction and fusion framework to achieve discriminative feature representations for gait recognition. Towards this goal, we take advantage of both global visual information and local region details and develop a Global and Local Feature Extractor (GLFE). Specifically, our GLFE module is composed of our newly designed multiple global and local convolutional layers (GLConv) to ensemble global and local features in a principle manner. Furthermore, we present a novel operation, namely Local Temporal Aggregation (LTA), to further preserve the spatial information by reducing the temporal resolution to obtain higher spatial resolution. With the help of our GLFE and LTA, our method significantly improves the discriminativeness of our visual features, thus improving the gait recognition performance. Extensive experiments demonstrate that our proposed method outperforms state-of-the-art gait recognition methods on two popular datasets.
****

Labels4Free: Unsupervised Segmentation Using StyleGAN
Rameen Abdal, Peihao Zhu, Niloy J. Mitra, Peter Wonka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13970-13979
We propose an unsupervised segmentation framework for StyleGAN generated objects. We build on two main observations. First, the features generated by StyleGAN hold valuable information that can be utilized towards training segmentation netw

orks. Second, the foreground and background can often be treated to be largely i
ndependent and be swapped across images to produce plausible composited images.
For our solution, we propose to augment the Style-GAN2 generator architecture wi
th a segmentation branch and to split the generator into a foreground and backgr
ound network. This enables us to generate soft segmentation masks for the foregr
ound object in an unsupervised fashion. On multiple object classes, we report co
mparable results against state-of-the-art supervised segmentation networks, whil
e against the best unsupervised segmentation approach we demonstrate a clear imp
rovement, both in qualitative and quantitative metrics. Project Page : https:/ra
meenabdal.github.io/Labels4Free
****

Harnessing the Conditioning Sensorium for Improved Image Translation
Cooper Nederhood, Nicholas Kolkin, Deqing Fu, Jason Salavon; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6752-6761
Existing methods for multi-modal domain translation learn to embed the input ima
ges into a domain-invariant "content" space and a domain-specific "style" space
from which novel images can be synthesized. Rather than learning to embed the RG
B image from scratch we propose deriving our content representation from conditi
oning data produced by pretrained off-the-shelf networks. Motivated by the inher
ent ambiguity of "content", which has different meanings depending on the desire
d level of abstraction, this approach gives intuitive control over which aspects
 of content are preserved across domains. We evaluate our method on traditional,
 well-aligned, datasets such as CelebA-HQ, and propose two novel datasets for ev
aluation on more complex scenes: ClassicTV and FFHQ-WildCrops. Our approach, whi
ch we call Sensorium, enables higher quality domain translation for complex scen
es than prior work.
****

DRINet: A Dual-Representation Iterative Learning Network for Point Cloud Segment
ation
Maosheng Ye, Shuangjie Xu, Tongyi Cao, Qifeng Chen; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2021, pp. 7447-7456
We present a novel and flexible architecture for point cloud segmentation with d
ual-representation iterative learning. In point cloud processing, different repr
esentations have their own pros and cons. Thus, finding suitable ways to represe
nt point cloud data structure while keeping its own internal physical property s
uch as permutation and scale-invariant is a fundamental problem. Therefore, we p
ropose our work, DRINet, which serves as the basic network structure for dual-re
presentation learning with great flexibility at feature transferring and less co
mputation cost, especially for large-scale point clouds. DRINet mainly consists
of two modules called Sparse Point-Voxel Feature Extraction and Sparse Voxel-Poi
nt Feature Extraction. By utilizing these two modules iteratively, features can
be propagated between two different representations. We further propose a novel
multi-scale pooling layer for pointwise locality learning to improve context inf
ormation propagation. Our network achieves state-of-the-art results for point cl
oud classification and segmentation tasks on several datasets while maintaining
high runtime efficiency. For large-scale outdoor scenarios, our method outperfor
ms state-of-the-art methods with a real-time inference speed of 62ms per frame.
****

Spectral Leakage and Rethinking the Kernel Size in CNNs
Nergis Tomen, Jan C. van Gemert; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2021, pp. 5138-5147
Convolutional layers in CNNs implement linear filters which decompose the input
into different frequency bands. However, most modern architectures neglect stand
ard principles of filter design when optimizing their model choices regarding th
e size and shape of the convolutional kernel. In this work, we consider the well
-known problem of spectral leakage caused by windowing artifacts in filtering op
erations in the context of CNNs. We show that the small size of CNN kernels make
 them susceptible to spectral leakage, which may induce performance-degrading ar
tifacts. To address this issue, we propose the use of larger kernel sizes along
with the Hamming window function to alleviate leakage in CNN architectures. We d

emonstrate improved classification accuracy on multiple benchmark datasets inclu
ding Fashion-MNIST, CIFAR-10, CIFAR-100 and ImageNet with the simple use of a st
andard window function in convolutional layers. Finally, we show that CNNs emplo
ying the Hamming window display increased robustness against various adversarial
 attacks.
****

High-Fidelity Pluralistic Image Completion With Transformers
Ziyu Wan, Jingbo Zhang, Dongdong Chen, Jing Liao; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2021, pp. 4692-4701
Image completion has made tremendous progress with convolutional neural networks
 (CNNs), because of their powerful texture modeling capacity. However, due to so
me inherent properties (eg, local inductive prior, spatial-invariant kernels), C
NNs do not perform well in understanding global structures or naturally support
pluralistic completion. Recently, transformers demonstrate their power in modeli
ng the long-term relationship and generating diverse results, but their computat
ion complexity is quadratic to input length, thus hampering the application in p
rocessing high-resolution images. This paper brings the best of both worlds to p
luralistic image completion: appearance prior reconstruction with transformer an
d texture replenishment with CNN. The former transformer recovers pluralistic co
herent structures together with some coarse textures, while the latter CNN enhan
ces the local texture details of coarse priors guided by the high-resolution mas
ked images. The proposed method vastly outperforms state-of-the-art methods in t
erms of three aspects: 1) large performance boost on image fidelity even compare
d to deterministic completion methods; 2) better diversity and higher fidelity f
or pluralistic completion; 3) exceptional generalization ability on large masks
and generic dataset, like ImageNet. Code and pre-trained models have been public
ly released at https://github.com/raywzy/ICT.
****

Dance With Self-Attention: A New Look of Conditional Random Fields on Anomaly De
tection in Videos
Didik Purwanto, Yie-Tarng Chen, Wen-Hsien Fang; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2021, pp. 173-183
This paper proposes a novel weakly supervised approach for anomaly detection, wh
ich begins with a relation-aware feature extractor to capture the multi-scale co
nvolutional neural network (CNN) features from a video. Afterwards, self-attenti
on is integrated with conditional random fields (CRFs), the core of the network,
 to make use of the ability of self-attention in capturing the short-range corre
lations of the features and the ability of CRFs in learning the inter-dependenci
es of these features. Such a framework can learn not only the spatio-temporal in
teractions among the actors which are important for detecting complex movements,
 but also their short- and long-term dependencies across frames. Also, to deal w
ith both local and non-local relationships of the features, a new variant of sel
f-attention is developed by taking into consideration a set of cliques with diff
erent temporal localities. Moreover, a contrastive multi-instance learning schem
e is considered to broaden the gap between the normal and abnormal instances, re
sulting in more accurate abnormal discrimination. Simulations reveal that the ne
w method provides superior performance to the state-of-the-art works on the wide
spread UCF-Crime and ShanghaiTech datasets.
****

Text Is Text, No Matter What: Unifying Text Recognition Using Knowledge Distilla
tion
Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Yi-Zhe Song; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
983-992
Text recognition remains a fundamental and extensively researched topic in compu
ter vision, largely owing to its wide array of commercial applications. The chal
lenging nature of the very problem however dictated a fragmentation of research
efforts: Scene Text Recognition (STR) that deals with text in everyday scenes, a
nd Handwriting Text Recognition (HTR) that tackles hand-written text. In this pa
per, for the first time, we argue for their unification -- we aim for a single m

odel that can compete favourably with two separate state-of-the-art STR and HTR models. We first show that cross-utilisation of STR and HTR models trigger significant performance drops due to differences in their inherent challenges. We then tackle their union by introducing a knowledge distillation (KD) based framework. This however is non-trivial, largely due to the variable-length and sequential nature of text sequences, which renders off-the-shelf KD techniques that mostly work with global fixed length data, inadequate. For that, we propose four distillation losses, all of which are specifically designed to cope with the aforementioned unique characteristics of text recognition. Empirical evidence suggests that our proposed unified model performs at par with individual models, even surpassing them in certain cases. Ablative studies demonstrate that naive baselines such as a two-stage framework, multi-task and domain adaption/generalisation alternatives do not work that well, further authenticating our design.

****

Unsupervised 3D Pose Estimation for Hierarchical Dance Video Recognition

Xiaodan Hu, Narendra Ahuja; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11015-11024

Dance experts often view dance as a hierarchy of information, spanning low-level (raw images, image sequences), mid-levels (human poses and bodypart movements), and high-level (dance genre). We propose a Hierarchical Dance Video Recognition framework (HDVR). HDVR estimates 2D pose sequences, tracks dancers, and then simultaneously estimates corresponding 3D poses and 3D-to-2D imaging parameters, without requiring ground truth for 3D poses. Unlike most methods that work on a single person, our tracking works on multiple dancers, under occlusions. From the estimated 3D pose sequence, HDVR extracts body part movements, and therefrom dance genre. The resulting hierarchical dance representation is explainable to experts. To overcome noise and interframe correspondence ambiguities, we enforce spatial and temporal motion smoothness and photometric continuity over time. We use an LSTM network to extract 3D movement subsequences from which we recognize dance genre. For experiments, we have identified 154 movement types, of 16 body parts, and assembled a new University of Illinois Dance (UID) Dataset, containing 1143 video clips of 9 genres covering 30 hours, annotated with movement and genre labels. Our experimental results demonstrate that our algorithms outperform the state-of-the-art 3D pose estimation methods, which also enhances our dance recognition performance.

****

What You Can Learn by Staring at a Blank Wall

Prafull Sharma, Miika Aittala, Yoav Y. Schechner, Antonio Torralba, Gregory W. Wornell, William T. Freeman, Frédo Durand; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2330-2339

We present a passive non-line-of-sight method that infers the number of people or activity of a person from the observation of a blank wall in an unknown room. Our technique analyzes complex imperceptible changes in indirect illumination in a video of the wall to reveal a signal that is correlated with motion in the hidden part of a scene. We use this signal to classify between zero, one, or two moving people, or the activity of a person in the hidden scene. We train two convolutional neural networks using data collected from 20 different scenes, and achieve an accuracy of approximately 94% for both tasks in unseen test environments and real-time online settings. Unlike other passive non-line-of-sight methods, the technique does not rely on known occluders or controllable light sources, and generalizes to unknown rooms with no recalibration. We analyze the generalization and robustness of our method with both real and synthetic data, and study the effect of the scene parameters on the signal quality.

****

Improving Contrastive Learning by Visualizing Feature Transformation

Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, Chang Wen Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10306-10315

Contrastive learning, which aims at minimizing the distance between positive pairs while maximizing that of negative ones, has been widely and successfully appl

ied in unsupervised feature learning, where the design of positive and negative (pos/neg) pairs is one of its keys. In this paper, we attempt to devise a feature-level data manipulation, differing from data augmentation, to enhance the generic contrastive self-supervised learning. To this end, we first design a visualization scheme for pos/neg score (pos/neg score indicates cosine similarity of pos/neg pair.) distribution, which enables us to analyze, interpret and understand the learning process. To our knowledge, this is the first attempt of its kind. More importantly, leveraging this tool, we gain some significant observations, which inspire our novel Feature Transformation proposals including the extrapolation of positives. This operation creates harder positives to boost the learning because hard positives enable the model to be more view-invariant. Besides, we propose the interpolation among negatives, which provides diversified negatives and makes the model more discriminative. It is the first attempt to deal with both challenges simultaneously. Experiment results show that our proposed Feature Transformation can improve at least 6.0% accuracy on ImageNet-100 over MoCo baseline, and about 2.0% accuracy on ImageNet-1K over the MoCoV2 baseline. Transferring to the downstream tasks successfully demonstrate our model is less task-bias. Visualization tools and codes: https://github.com/DTennant/CL-Visualizing-Feature-Transformation.

****

Complementary Patch for Weakly Supervised Semantic Segmentation

Fei Zhang, Chaochen Gu, Chenyue Zhang, Yuchao Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7242-7251

Weakly Supervised Semantic Segmentation (WSSS) based on image-level labels has been greatly advanced by exploiting the outputs of Class Activation Map (CAM) to generate the pseudo labels for semantic segmentation. However, CAM merely discovers seeds from a small number of regions, which may be insufficient to serve as pseudo masks for semantic segmentation. In this paper, we formulate the expansion of object regions in CAM as an increase in information. From the perspective of information theory, we propose a novel Complementary Patch (CP) Representation and prove that the information of the sum of the CAMs by a pair of input images with complementary hidden (patched) parts, namely CP Pair, is greater than or equal to the information of the baseline CAM. Therefore, a CAM with more information related to object seeds can be obtained by narrowing down the gap between the sum of CAMs generated by the CP Pair and the original CAM. We propose a CP Network (CPN) implemented by a triplet network and three regularization functions. To further improve the quality of the CAMs, we propose a Pixel-Region Correlation Module (PRCM) to augment the contextual information by using object-region relations between the feature maps and the CAMs. Experimental results on the PASCAL VOC 2012 datasets show that our proposed method achieves a new state-of-the-art in WSSS, validating the effectiveness of our CP Representation and CPN.

****

Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining

Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, Xiaodan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11782-11791

Nowadays, customer's demands for E-commerce are more diversified, which introduces more complications to the product retrieval industry. Previous methods are either subject to single-modal input or perform supervised image-level product retrieval, thus fail to accommodate real-life scenarios where enormous weakly annotated multi-modal data are present. In this paper, we investigate a more realistic setting that aims to perform weakly-supervised multi-modal instance-level product retrieval among fine-grained product categories. To promote the study of this challenging task, we contribute Product1M, one of the largest multi-modal cosmetic datasets for real-world instance-level retrieval. Notably, Product1M contains over 1 million image-caption pairs and consists of two sample types, i.e., single-product and multi-product samples, which encompass a wide variety of cosmetics brands. In addition to the great diversity, Product1M enjoys several appealing characteristics including fine-grained categories, complex combinations, and

fuzzy correspondence that well mimic the real-world scenes. Moreover, we propose a novel model named Cross-modal contrAstive Product Transformer for instance-level prodUct REtrieval (CAPTURE), that excels in capturing the potential synergy between multi-modal inputs via a hybrid-stream transformer in a self-supervised manner. CAPTURE generates discriminative instance features via masked multi-modal learning as well as cross-modal contrastive pretraining and it outperforms several SOTA cross-modal baselines. Extensive ablation studies well demonstrate the effectiveness and the generalization capacity of our model.

****

StyleFormer: Real-Time Arbitrary Style Transfer via Parametric Style Composition
Xiaolei Wu, Zhihao Hu, Lu Sheng, Dong Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14618-14627
In this work, we propose a new feed-forward arbitrary style transfer method, referred to as StyleFormer, which can simultaneously fulfill fine-grained style diversity and semantic content coherency. Specifically, our transformer-inspired feature-level stylization method consists of three modules: (a) the style bank generation module for sparse but compact parametric style pattern extraction, (b) the transformer-driven style composition module for content-guided global style composition, and (c) the parametric content modulation module for flexible but faithful stylization. The output stylized images are impressively coherent with the content structure, sensitive to the detailed style variations, but still holistically adhere to the style distributions from the style images. Qualitative and quantitative comparisons as well as comprehensive user studies demonstrate that our StyleFormer outperforms the existing SOTA methods in generating visually plausible stylization results with real-time efficiency.

****

Hypercorrelation Squeeze for Few-Shot Segmentation
Juhong Min, Dahyun Kang, Minsu Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6941-6952
Few-shot semantic segmentation aims at learning to segment a target object from a query image using only a few annotated support images of the target class. This challenging task requires to understand diverse levels of visual cues and analyze fine-grained correspondence relations between the query and the support images. To address the problem, we propose Hypercorrelation Squeeze Networks (HSNet) that leverages multi-level feature correlation and efficient 4D convolutions. It extracts diverse features from different levels of intermediate convolutional layers and constructs a collection of 4D correlation tensors, i.e., hypercorrelations. Using efficient center-pivot 4D convolutions in a pyramidal architecture, the method gradually squeezes high-level semantic and low-level geometric cues of the hypercorrelation into precise segmentation masks in coarse-to-fine manner. The significant performance improvements on standard few-shot segmentation benchmarks of PASCAL-5i, COCO-20i, and FSS-1000 verify the efficacy of the proposed method.

****

Digging Into Uncertainty in Self-Supervised Multi-View Stereo
Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6078-6087
Self-supervised Multi-view stereo (MVS) with a pretext task of image reconstruction has achieved significant progress recently. However, previous methods are built upon intuitions, lacking comprehensive explanations about the effectiveness of the pretext task in self-supervised MVS. To this end, we propose to estimate epistemic uncertainty in self-supervised MVS, accounting for what the model ignores. Specially, the limitations can be resorted into two folds: ambiguious supervision in foreground and noisy disturbance in background. To address these issues, we propose a novel Uncertainty reduction Multi-view Stereo (U-MVS) framework for self-supervised learning. To alleviate ambiguous supervision in foreground, we involve extra correspondence prior with a flow-depth consistency loss. The dense 2D correspondence of optical flows is used to regularize the 3D stereo correspondence in MVS. To handle the noisy disturbance in background, we use Monte-Ca

rlo Dropout to acquire the uncertainty map and further filter the unreliable sup ervision signals on invalid regions. Extensive experiments on DTU and Tank&Templ es benchmark show that our U-MVS framework achieves the best performance among u nsupervised MVS methods, with competitive performance with its supervised oppone nts.
****

A Style and Semantic Memory Mechanism for Domain Generalization
Yang Chen, Yu Wang, Yingwei Pan, Ting Yao, Xinmei Tian, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9164-9173
Mainstream state-of-the-art domain generalization algorithms tend to prioritize the assumption on semantic invariance across domains. Meanwhile, the inherent in tra-domain style invariance is usually underappreciated and put on the shelf. In this paper, we reveal that leveraging intra-domain style invariance is also of pivotal importance in improving the efficiency of domain generalization. We veri fy that it is critical for the network to be informative on what domain features are invariant and shared among instances, so that the network sharpens its unde rstanding and improves its semantic discriminative ability. Correspondingly, we also propose a novel "jury" mechanism, which is particularly effective in learni ng useful semantic feature commonalities among domains. Our complete model calle d STEAM can be interpreted as a novel probabilistic graphical model, for which t he implementation requires convenient constructions of two kinds of memory banks : semantic feature bank and style feature bank. Empirical results show that our proposed framework surpasses the state-of-the-art methods by clear margins.
****

EigenGAN: Layer-Wise Eigen-Learning for GANs
Zhenliang He, Meina Kan, Shiguang Shan; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2021, pp. 14408-14417
Recent studies on Generative Adversarial Network (GAN) reveal that different lay ers of a generative CNN hold different semantics of the synthesized images. Howe ver, few GAN models have explicit dimensions to control the semantic attributes represented in a specific layer. This paper proposes EigenGAN which is able to u nsupervisedly mine interpretable and controllable dimensions from different gene rator layers. Specifically, EigenGAN embeds one linear subspace with orthogonal basis into each generator layer. Via generative adversarial training to learn a target distribution, these layer-wise subspaces automatically discover a set of "eigen-dimensions" at each layer corresponding to a set of semantic attributes o r interpretable variations. By traversing the coefficient of a specific eigen-di mension, the generator can produce samples with continuous changes corresponding to a specific semantic attribute. Taking the human face for example, EigenGAN c an discover controllable dimensions for high-level concepts such as pose and gen der in the subspace of deep layers, as well as low-level concepts such as hue an d color in the subspace of shallow layers. Moreover, in the linear case, we theo retically prove that our algorithm derives the principal components as PCA does. Codes can be found in https://github.com/LynnHo/EigenGAN-Tensorflow.
****

Uncertainty-Aware Human Mesh Recovery From Video by Learning Part-Based 3D Dynam ics
Gun-Hee Lee, Seong-Whan Lee; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2021, pp. 12375-12384
Despite the recent success of 3D human reconstruction methods, recovering the ac curate and smooth 3D human motion from video is still challenging. Designing a t emporal model in the encoding stage is not sufficient enough to settle the trade -off problem between the per-frame accuracy and the motion smoothness. To addres s this problem, we approach some of the fundamental problems of 3D reconstructio n tasks, simultaneously predicting 3D pose and 3D motion dynamics. First, we uti lize the power of uncertainty to address the problem of multiple 3D configuratio ns resulting in the same 2D projections. Second, we confirmed that dividing the body into local regions shows outstanding results for estimating 3D motion dynam ics. In this paper, we propose (i) an encoder that makes two different estimatio

ns: a static feature that presents 2D pose feature as distribution and a dynamic feature that includes optical flow information and (ii) a decoder that divides the body into five different local regions to estimate the 3D motion dynamics of each region. We demonstrate how our method recovers the accurate and smooth motion and achieves the state-of-the-art results for both constrained and in-the-wild videos.
****

Neural TMDlayer: Modeling Instantaneous Flow of Features via SDE Generators
Zihang Meng, Vikas Singh, Sathya N. Ravi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11635-11644
We study how stochastic differential equation (SDE) based ideas can inspire new modifications to existing algorithms for a set of problems in computer vision. Loosely speaking, our formulation is related to both explicit and implicit strategies for data augmentation and group equivariance, but is derived from new results in the SDE literature on estimating infinitesimal generators of a class of stochastic processes. If and when there is nominal agreement between the needs of an application/task and the inherent properties and behavior of the types of processes that we can efficiently handle, we obtain a very simple and efficient plug-in layer that can be incorporated within any existing network architecture, with minimal modification and only a few additional parameters. We show promising experiments on a number of vision tasks including few shot learning, point cloud transformers and deep variational segmentation obtaining efficiency or performance improvements.
****

Rethinking Transformer-Based Set Prediction for Object Detection
Zhiqing Sun, Shengcao Cao, Yiming Yang, Kris M. Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3611-3620
DETR is a recently proposed Transformer-based method which views object detection as a set prediction problem and achieves state-of-the-art performance but demands extra-long training time to converge. In this paper, we investigate the causes of the optimization difficulty in the training of DETR. Our examinations reveal several factors contributing to the slow convergence of DETR, primarily the issues with the Hungarian loss and the Transformer cross attention mechanism. To overcome these issues we propose two solutions, namely, TSP-FCOS (Transformer-based Set Prediction with FCOS) and TSP-RCNN (Transformer-based Set Prediction with RCNN). Experimental results show that the proposed methods not only converge much faster than the original DETR, but also significantly outperform DETR and other baselines in terms of detection accuracy.
****

FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras
Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, Alex Kendall; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15273-15282
Driving requires interacting with road agents and predicting their future behaviour in order to navigate safely. We present FIERY: a probabilistic future prediction model in bird's-eye view from monocular cameras. Our model predicts future instance segmentation and motion of dynamic agents that can be transformed into non-parametric future trajectories. Our approach combines the perception, sensor fusion and prediction components of a traditional autonomous driving stack by estimating bird's-eye-view prediction directly from surround RGB monocular camera inputs. FIERY learns to model the inherent stochastic nature of the future solely from camera driving data in an end-to-end manner, without relying on HD maps, and predicts multimodal future trajectories. We show that our model outperforms previous prediction baselines on the NuScenes and Lyft datasets. The code and trained models are available at https://github.com/wayveai/fiery.
****

CLEAR: Clean-Up Sample-Targeted Backdoor in Neural Networks
Liuwan Zhu, Rui Ning, Chunsheng Xin, Chonggang Wang, Hongyi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16453-

16462
The data poisoning attack has raised serious security concerns on the safety of deep neural networks since it can lead to neural backdoor that misclassifies certain inputs crafted by an attacker. In particular, the sample-targeted backdoor attack is a new challenge. It targets at one or a few specific samples, called target samples, to misclassify them to a target class. Without a trigger planted in the backdoor model, the existing backdoor detection schemes fail to detect the sample-targeted backdoor as they depend on reverse-engineering the trigger or strong features of the trigger. In this paper, we propose a novel scheme to detect and mitigate sample-targeted backdoor attacks. We discover and demonstrate a unique property of the sample-targeted backdoor, which forces a boundary change such that small "pockets" are formed around the target sample. Based on this observation, we propose a novel defense mechanism to pinpoint a malicious pocket by "wrapping" them into a tight convex hull in the feature space. We design an effective algorithm to search for such a convex hull and remove the backdoor by fine-tuning the model using the identified malicious samples with the corrected label according to the convex hull. The experiments show that the proposed approach is highly efficient for detecting and mitigating a wide range of sample-targeted backdoor attacks.
****

Motion-Augmented Self-Training for Video Recognition at Smaller Scale
Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, Cees G. M. Snoek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10429-10438
The goal of this paper is to self-train a 3D convolutional neural network on an unlabeled video collection for deployment on small-scale video collections. As smaller video datasets benefit more from motion than appearance, we strive to train our network using optical flow, but avoid its computation during inference. We propose the first motion-augmented self-training regime, we call MotionFit. We start with supervised training of a motion model on a small, and labeled, video collection. With the motion model we generate pseudo-labels for a large unlabeled video collection, which enables us to transfer knowledge by learning to predict these pseudo-labels with an appearance model. Moreover, we introduce a multi-clip loss as a simple yet efficient way to improve the quality of the pseudo-labeling, even without additional auxiliary tasks. We also take into consideration the temporal granularity of videos during self-training of the appearance model, which was missed in previous works. As a result we obtain a strong motion-augmented representation model suited for video downstream tasks like action recognition and clip retrieval. On small-scale video datasets, MotionFit outperforms alternatives for knowledge transfer by 5%-8%, video-only self-supervision by 1%-7% and semisupervised learning by 9%-18% using the same amount of class labels.
****

Generating Smooth Pose Sequences for Diverse Human Motion Prediction
Wei Mao, Miaomiao Liu, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13309-13318
Recent progress in stochastic motion prediction, i.e., predicting multiple possible future human motions given a single past pose sequence, has led to producing truly diverse future motions and even providing control over the motion of some body parts. However, to achieve this, the state-of-the-art method requires learning several mappings for diversity and a dedicated model for controllable motion prediction. In this paper, we introduce a unified deep generative network for both diverse and controllable motion prediction. To this end, we leverage the intuition that realistic human motions consist of smooth sequences of valid poses, and that, given limited data, learning a pose prior is much more tractable than a motion one. We therefore design a generator that predicts the motion of different body parts sequentially, and introduce a normalizing flow based pose prior, together with a joint angle loss, to achieve motion realism.Our experiments on two standard benchmark datasets, Human3.6M and HumanEva-I, demonstrate that our approach outperforms the state-of-the-art baselines in terms of both sample diversity and accuracy. The code is available at https://github.com/wei-mao-2019/gsp

s
****
DensePose 3D: Lifting Canonical Surface Maps of Articulated Objects to the Third Dimension
Roman Shapovalov, David Novotny, Benjamin Graham, Patrick Labatut, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11729-11739
We tackle the problem of monocular 3D reconstruction of articulated objects like humans and animals. Our key contribution is DensePose 3D, a novel parametric model of an articulated mesh, which can be learned in a self-supervised fashion from 2D image annotations only. This is in stark contrast with previous human body reconstruction methods that utilize a parametric model like SMPL pre-trained on a large dataset of 3D body scans that had to be obtained in a controlled environment. DensePose 3D can thus be applied for modelling broad range of articulated categories such as animal species. In an end-to-end fashion, it automatically learns to softly assign each vertex of a category-specific 3D template mesh to one of the rigidly moving latent parts and trains a single-view network predicting rigid motions of the parts to deform the template so that it re-projects correctly to the dense 2D surface annotations of objects (such as DensePose). In order to prevent unrealistic template deformations, we further propose to align the motions of nearby mesh vertices by expressing the part assignment as a function of the smooth eigenfunctions of the Laplace--Beltrami operator computed on the template mesh. Our experiments demonstrate improvements over the state-of-the-art non-rigid structure-from-motion baselines on both synthetic and real data on categories of humans and animals.
****
Unpaired Learning for Deep Image Deraining With Rain Direction Regularizer
Yang Liu, Ziyu Yue, Jinshan Pan, Zhixun Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4753-4761
We present a simple yet effective unpaired learning based image rain removal method from an unpaired set of synthetic images and real rainy images by exploring the properties of rain maps. The proposed algorithm mainly consists of a semi-supervised learning part and a knowledge distillation part. The semi-supervised part estimates the rain map and reconstructs the derained image based on the well-established layer separation principle. To facilitate rain removal, we develop a rain direction regularizer to constrain the rain estimation network in the semi-supervised learning part. With the estimated rain maps from the semi-supervised learning part, we first synthesize a new paired set by adding to rain-free images based on the superimposition model. The real rainy images and the derained results constitute another paired set. Then we develop an effective knowledge distillation method to explore such two paired sets so that the deraining model in the semi-supervised learning part is distilled. We propose two new rainy datasets, named RainDirection and Real3000, to validate the effectiveness of the proposed method. Both quantitative and qualitative experimental results demonstrate that the proposed method achieves favorable results against state-of-the-art methods in benchmark datasets and real-world images.
****
Self-Supervised Image Prior Learning With GMM From a Single Noisy Image
Haosen Liu, Xuan Liu, Jiangbo Lu, Shan Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2845-2854
The lack of clean images undermines the practicability of supervised image prior learning methods, of which the training schemes require a large number of clean images. To free image prior learning from the image collection burden, a novel Self-Supervised learning method for Gaussian Mixture Model (SS-GMM) is proposed in this paper. It can simultaneously achieve the noise level estimation and the image prior learning directly from only a single noisy image. This work is derived from our study on eigenvalues of the GMM's covariance matrix. Through statistical experiments and theoretical analysis, we conclude that (1) covariance eigenvalues for clean images hold the sparsity; and that (2) those for noisy images contain sufficient information for noise estimation. The first conclusion inspire

s us to impose a sparsity constraint on covariance eigenvalues during the learni ng process to suppress the influence of noise. The second conclusion leads to a self-contained noise estimation module of high accuracy in our proposed method. This module serves to estimate the noise level and automatically determine the s pecific level of the sparsity constraint. Our final derived method requires only minor modifications to the standard expectation-maximization algorithm. This ma kes it easy to implement. Very interestingly, the GMM learned via our proposed s elf-supervised learning method can even achieve better image denoising performan ce than its supervised counterpart, i.e., the EPLL. Also, it is on par with the state-of-the-art self-supervised deep learning method, i.e., the Self2Self. Code is available at https://github.com/HUST-Tan/SS-GMM.
****

Exploiting a Joint Embedding Space for Generalized Zero-Shot Semantic Segmentati on

Donghyeon Baek, Youngmin Oh, Bumsub Ham; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2021, pp. 9536-9545

We address the problem of generalized zero-shot semantic segmentation (GZS3) pre dicting pixel-wise semantic labels for seen and unseen classes. Most GZS3 method s adopt a generative approach that synthesizes visual features of unseen classes from corresponding semantic ones (e.g., word2vec) to train novel classifiers fo r both seen and unseen classes. Although generative methods show decent performa nce, they have two limitations: (1) the visual features are biased towards seen classes; (2) the classifier should be retrained whenever novel unseen classes ap pear. We propose a discriminative approach to address these limitations in a uni fied framework. To this end, we leverage visual and semantic encoders to learn a joint embedding space, where the semantic encoder transforms semantic features to semantic prototypes that act as centers for visual features of corresponding classes. Specifically, we introduce boundary-aware regression (BAR) and semantic consistency (SC) losses to learn discriminative features. Our approach to explo iting the joint embedding space, together with BAR and SC terms, alleviates the seen bias problem. At test time, we avoid the retraining process by exploiting s emantic prototypes as a nearest-neighbor (NN) classifier. To further alleviate t he bias problem, we also propose an inference technique, dubbed Apollonius calib ration (AC), that modulates the decision boundary of the NN classifier to the Ap ollonius circle adaptively. Experimental results demonstrate the effectiveness o f our framework, achieving a new state of the art on standard benchmarks.
****

HeadGAN: One-Shot Neural Head Synthesis and Editing

Michail Christos Doukas, Stefanos Zafeiriou, Viktoriia Sharmanska; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 143 98-14407

Recent attempts to solve the problem of head reenactment using a single referenc e image have shown promising results. However, most of them either perform poorl y in terms of photo-realism, or fail to meet the identity preservation problem, or do not fully transfer the driving pose and expression. We propose HeadGAN, a novel system that conditions synthesis on 3D face representations, which can be extracted from any driving video and adapted to the facial geometry of any refer ence image, disentangling identity from expression. We further improve mouth mov ements, by utilising audio features as a complementary input. The 3D face repres entation enables HeadGAN to be further used as an efficient method for compressi on and reconstruction and a tool for expression and pose editing.
****

Aligning Subtitles in Sign Language Videos

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, A ndrew Zisserman; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2021, pp. 11552-11561

The goal of this work is to temporally align asynchronous subtitles in sign lang uage videos. In particular, we focus on sign-language interpreted TV broadcast d ata comprising (i) a video of continuous signing, and (ii) subtitles correspondi ng to the audio content. Previous work exploiting such weakly-aligned data only

considered finding keyword-sign correspondences, whereas we aim to localise a co
mplete subtitle text in continuous signing. We propose a Transformer architectur
e tailored for this task, which we train on manually annotated alignments coveri
ng over 15K subtitles that span 17.7 hours of video. We use BERT subtitle embedd
ings and CNN video representations learned for sign recognition to encode the tw
o signals, which interact through a series of attention layers. Our model output
s frame-level predictions, i.e., for each video frame, whether it belongs to the
 queried subtitle or not. Through extensive evaluations, we show substantial imp
rovements over existing alignment baselines that do not make use of subtitle tex
t embeddings for learning. Our automatic alignment model opens up possibilities
for advancing machine translation of sign languages via providing continuously s
ynchronized video-text data.
****

Variational Feature Disentangling for Fine-Grained Few-Shot Classification
Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, Dimitris Samaras; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
8812-8821
Data augmentation is an intuitive step towards solving the problem of few-shot c
lassification. However, ensuring both discriminability and diversity in the augm
ented samples is challenging. To address this, we propose a feature disentanglem
ent framework that allows us to augment features with randomly sampled intra-cla
ss variations while preserving their class-discriminative features. Specifically
, we disentangle a feature representation into two components: one represents th
e intra-class variance and the other encodes the class-discriminative informatio
n. We assume that the intra-class variance induced by variations in poses, backg
rounds, or illumination conditions is shared across all classes and can be model
led via a common distribution. Then we sample features repeatedly from the learn
ed intra-class variability distribution and add them to the class-discriminative
 features to get the augmented features. Such a data augmentation scheme ensures
 that the augmented features inherit crucial class-discriminative features while
 exhibiting large intra-class variance. Our method significantly outperforms the
 state-of-the-art methods on multiple challenging fine-grained few-shot image cl
assification benchmarks. Code is available at: https://github.com/cvlab-stonybro
ok/vfd-iccv21
****

MultiSiam: Self-Supervised Multi-Instance Siamese Representation Learning for Au
tonomous Driving
Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7546-7554
Autonomous driving has attracted much attention over the years but turns out to
be harder than expected, probably due to the difficulty of labeled data collecti
on for model training. Self-supervised learning (SSL), which leverages unlabeled
 data only for representation learning, might be a promising way to improve mode
l performance. Existing SSL methods, however, usually rely on the single-centric
-object guarantee, which may not be applicable for multi-instance datasets such
as street scenes. To alleviate this limitation, we raise two issues to solve: (1
) how to define positive samples for cross-view consistency and (2) how to measu
re similarity in multi-instance circumstances. We first adopt an IoU threshold d
uring random cropping to transfer global-inconsistency to local-consistency. The
n, we propose two feature alignment methods to enable 2D feature maps for multi-
instance similarity measurement. Additionally, we adopt intra-image clustering w
ith self-attention for further mining intra-image similarity and translation-inv
ariance. Experiments show that, when pre-trained on Waymo dataset, our method ca
lled Multi-instance Siamese Network (MultiSiam) remarkably improves generalizati
on ability and achieves state-of-the-art transfer performance on autonomous driv
ing benchmarks, including Cityscapes and BDD100K, while existing SSL counterpart
s like MoCo, MoCo-v2, and BYOL show significant performance drop. By pre-trainin
g on SODA10M, a large-scale autonomous driving dataset, MultiSiam exceeds the Im
ageNet pre-trained MoCo-v2, demonstrating the potential of domain-specific pre-t
raining. Code will be available at https://github.com/KaiChen1998/MultiSiam .

****
Pano-AVQA: Grounded Audio-Visual Question Answering on 360deg Videos
Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, Gunhee Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2031-2041
360deg videos convey holistic views for the surroundings of a scene. It provides audio-visual cues beyond predetermined normal field of views and displays distinctive spatial relations on a sphere. However, previous benchmark tasks for panoramic videos are still limited to evaluate the semantic understanding of audio-visual relationships or spherical spatial property in surroundings. We propose a novel benchmark named Pano-AVQA as a large-scale grounded audio-visual question answering dataset on panoramic videos. Using 5.4K 360deg video clips harvested online, we collect two types of novel question-answer pairs with bounding-box grounding: spherical spatial relation QAs and audio-visual relation QAs. We train several transformer-based models from Pano-AVQA, where the results suggest that our proposed spherical spatial embeddings and multimodal training objectives fairly contribute to better semantic understanding of the panoramic surroundings on the dataset.
****
Deep Implicit Surface Point Prediction Networks
Rahul Venkatesh, Tejan Karmali, Sarthak Sharma, Aurobrata Ghosh, R. Venkatesh Babu, László A. Jeni, Maneesh Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12653-12662
Deep neural representations of 3D shapes as implicit functions have been shown to produce high fidelity models surpassing the resolution-memory trade-off faced by the explicit representations using meshes and point clouds. However, most such approaches focus on representing closed shapes. Unsigned distance function (UDF) based approaches have been proposed recently as a promising alternative to represent both open and closed shapes. However, since the gradients of UDFs vanish on the surface, it is challenging to estimate local (differential) geometric properties like the normals and tangent planes which are needed for many downstream applications in vision and graphics. There are additional challenges in computing these properties efficiently with a low-memory footprint. This paper presents a novel approach that models such surfaces using a new class of implicit representations called the closest surface-point CSP representation. We show that CSP allows us to represent complex surfaces of any topology (open or closed) with high fidelity. It also allows for accurate and efficient computation of local geometric properties. We further demonstrate that it leads to efficient implementation of downstream algorithms like sphere-tracing for rendering the 3D surface as well as to create explicit mesh-based representations. Extensive experimental evaluation on the ShapeNet dataset validate the above contributions with results surpassing the state-of-the-art. Code and data are available at https://sites.google.com/view/cspnet
****
Broaden Your Views for Self-Supervised Video Learning
Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica P■tr■ucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1255-1265
Most successful self-supervised learning methods are trained to align the representations of two independent views from the data. State-of-the-art methods in video are inspired by image techniques, where these two views are similarly extracted by cropping and augmenting the resulting crop. However, these methods miss a crucial element in the video domain: time. We introduce BraVe, a self-supervised learning framework for video. In BraVe, one of the views has access to a narrow temporal window of the video while the other view has a broad access to the video content. Our models learn to generalise from the narrow view to the general content of the video. Furthermore, BraVe processes the views with different backbones, enabling the use of alternative augmentations or modalities into the broa

d view such as optical flow, randomly convolved RGB frames, audio or their combi nations. We demonstrate that BraVe achieves state-of-the-art results in self-sup ervised representation learning on standard video and audio classification bench marks including UCF101, HMDB51, Kinetics, ESC-50 and AudioSet.
****

Deep Metric Learning for Open World Semantic Segmentation
Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, Ming Liu; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15333-15342
Classical close-set semantic segmentation networks have limited ability to detec t out-of-distribution (OOD) objects, which is important for safety-critical appl ications such as autonomous driving. Incrementally learning these OOD objects wi th few annotations is an ideal way to enlarge the knowledge base of the deep lea rning models. In this paper, we propose an open world semantic segmentation syst em that includes two modules: (1) an open-set semantic segmentation module to de tect both in-distribution and OOD objects. (2) an incremental few-shot learning module to gradually incorporate those OOD objects into its existing knowledge ba se. This open world semantic segmentation system behaves like a human being, whi ch is able to identify OOD objects and gradually learn them with corresponding s upervision. We adopt the Deep Metric Learning Network (DMLNet) with contrastive clustering to implement open-set semantic segmentation. Compared to other open-s et semantic segmentation methods, our DMLNet achieves state-of-the-art performan ce on three challenging open-set semantic segmentation datasets without using ad ditional data or generative models. On this basis, two incremental few-shot lear ning methods are further proposed to progressively improve the DMLNet with the a nnotations of OOD objects.
****

Boundary-Sensitive Pre-Training for Temporal Localization in Videos
Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martínez, Xiatian Zhu , Li Zhang, Bernard Ghanem, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7220-7230
Many video analysis tasks require temporal localization for the detection of con tent changes. However, most existing models developed for these tasks are pre-tr ained on general video action classification tasks. This is due to large scale a nnotation of temporal boundaries in untrimmed videos being expensive. Therefore, no suitable datasets exist that enable pre-training in a manner sensitive to te mporal boundaries. In this paper for the first time, we investigate model pre-tr aining for temporal localization by introducing a novel boundary-sensitive prete xt (BSP) task. Instead of relying on costly manual annotations of temporal bound aries, we propose to synthesize temporal boundaries in existing video action cla ssification datasets. By defining different ways of synthesizing boundaries, BSP can then be simply conducted in a self-supervised manner via the classification of the boundary types. This enables the learning of video representations that are much more transferable to downstream temporal localization tasks. Extensive experiments show that the proposed BSP is superior and complementary to the exis ting action classification-based pre-training counterpart, and achieves new stat e-of-the-art performance on several temporal localization tasks. Please visit ou r website for more details https://frostinassiky.github.io/bsp.
****

SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation
Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12396-12405
Directly regressing all 6 degrees-of-freedom (6DoF) for the object pose (i.e. th e 3D rotation and translation) in a cluttered environment from a single RGB imag e is a challenging problem. While end-to-end methods have recently demonstrated promising results at high efficiency, they are still inferior when compared with elaborate PnP/RANSAC-based approaches in terms of pose accuracy. In this work, we address this shortcoming by means of a novel reason-ing about self-occlusion, in order to establish a two-layer representation for 3D objects which considera bly enhances the accuracy of end-to-end 6D pose estimation. Our frame-work, name

d SO-Pose, takes a single RGB image as input and respectively generates 2D-3D co rrespondences as well as self-occlusion information harnessing a shared encoder and two separate decoders. Both outputs are then fused to directly regress the 6 DoF pose parameters. Incorporating cross-layer consistencies that align correspo ndences, self-occlusion, and 6D pose, we can further improve accuracy and robust ness, surpassing or rivaling all other state-of-the-art approaches on various ch allenging datasets.
****

Explainable Video Entailment With Grounded Visual Evidence
Junwen Chen, Yu Kong; Proceedings of the IEEE/CVF International Conference on Co mputer Vision (ICCV), 2021, pp. 2021-2030
Video entailment aims at determining if a hypothesis textual statement is entail ed or contradicted by a premise video. The main challenge of video entailment is that it requires fine-grained reasoning to understand the complex and long stor y-based videos. To this end, we propose to incorporate visual grounding to the e ntailment by explicitly linking the entities described in the statement to the e vidence in the video. If the entities are grounded in the video, we enhance the entailment judgment by focusing on the frames where the entities occur. Besides, in entailment dataset, the real/fake statements are formed in pairs with subtle discrepancy, which allows an add-on explanation module to predict which words o r phrases make the statement contradictory to the video and regularize the train ing of the entailment judgment. Experimental results demonstrate that our approa ch significantly outperforms the state-of-the-art methods.
****

HiT: Hierarchical Transformer With Momentum Contrast for Video-Text Retrieval
Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, Zhongyuan Wang; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 11915-11925
Video-Text Retrieval has been a hot research topic with the growth of multimedia data on the internet. Transformer for video-text learning has attracted increas ing attention due to its promising performance. However, existing cross-modal tr ansformer approaches typically suffer from two major limitations: 1) Exploitatio n of the transformer architecture where different layers have different feature characteristics is limited; 2) End-to-end training mechanism limits negative sam ple interactions in a mini-batch. In this paper, we propose a novel approach nam ed Hierarchical Transformer (HiT) for video-text retrieval. HiT performs Hierarc hical Cross-modal Contrastive Matching in both feature-level and semantic-level, achieving multi-view and comprehensive retrieval results. Moreover, inspired by MoCo, we propose Momentum Cross-modal Contrast for cross-modal learning to enab le large-scale negative sample interactions on-the-fly, which contributes to the generation of more precise and discriminative representations. Experimental res ults on the three major Video-Text Retrieval benchmark datasets demonstrate the advantages of our method.
****

Unsupervised Few-Shot Action Recognition via Action-Appearance Aligned Meta-Adap tation
Jay Patravali, Gaurav Mittal, Ye Yu, Fuxin Li, Mei Chen; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 8484-8494
We present MetaUVFS as the first Unsupervised Meta-learning algorithm for Video Few-Shot action recognition. MetaUVFS leverages over 550K unlabeled videos to tr ain a two-stream 2D and 3D CNN architecture via contrastive learning to capture the appearance-specific spatial and action-specific spatio-temporal video featur es respectively. MetaUVFS comprises a novel Action-Appearance Aligned Meta-adapt ation (A3M) module that learns to focus on the action-oriented video features in relation to the appearance features via explicit few-shot episodic meta-learnin g over unsupervised hard-mined episodes. Our action-appearance alignment and exp licit few-shot learner conditions the unsupervised training to mimic the downstr eam few-shot task, enabling MetaUVFS to significantly outperform all unsupervise d methods on few-shot benchmarks. Moreover, unlike previous few-shot action reco gnition methods that are supervised, MetaUVFS needs neither base-class labels no

r a supervised pretrained backbone. Thus, we need to train MetaUVFS just once to perform competitively or sometimes even outperform state-of-the-art supervised methods on popular HMDB51, UCF101, and Kinetics100 few-shot datasets.
****

Poly-NL: Linear Complexity Non-Local Layers With 3rd Order Polynomials
Francesca Babiloni, Ioannis Marras, Filippos Kokkinos, Jiankang Deng, Grigorios Chrysos, Stefanos Zafeiriou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10518-10528
Spatial self-attention layers, in the form of Non-Local blocks, introduce long-range dependencies in Convolutional Neural Networks by computing pairwise similarities among all possible positions. Such pairwise functions underpin the effectiveness of non-local layers, but also determine a complexity that scales quadratically with respect to the input size both in space and time. This is a severely limiting factor that practically hinders the applicability of non-local blocks to even moderately sized inputs. Previous works focused on reducing the complexity by modifying the underlying matrix operations, however in this work we aim to retain full expressiveness of non-local layers while keeping complexity linear. We overcome the efficiency limitation of non-local blocks by framing them as special cases of 3rd order polynomial functions. This fact enables us to formulate novel fast Non-Local blocks, capable of reducing the complexity from quadratic to linear with no loss in performance, by replacing any direct computation of pairwise similarities with element-wise multiplications. The proposed method, which we dub as "Poly-NL", is competitive with state-of-the-art performance across image recognition, instance segmentation, and face detection tasks, while having considerably less computational overhead.
****

PatchMatch-RL: Deep MVS With Pixelwise Depth, Normal, and Visibility
Jae Yong Lee, Joseph DeGol, Chuhang Zou, Derek Hoiem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6158-6167
Recent learning-based multi-view stereo (MVS) methods show excellent performance with dense cameras and small depth ranges. However, non-learning based approaches still outperform for scenes with large depth ranges and sparser wide-baseline views, in part due to their PatchMatch optimization over pixelwise estimates of depth, normals, and visibility. In this paper, we propose an end-to-end trainable PatchMatch-based MVS approach that combines advantages of trainable costs and regularizations with pixelwise estimates. To overcome the challenge of the non-differentiable PatchMatch optimization that involves iterative sampling and hard decisions, we use reinforcement learning to minimize expected photometric cost and maximize likelihood of ground truth depth and normals. We incorporate normal estimation by using dilated patch kernels, and propose a recurrent cost regularization that applies beyond frontal plane-sweep algorithms to our pixelwise depth/normal estimates. We evaluate our method on widely used MVS benchmarks, ETH3D and Tanks and Temples (TnT), and compare to other state of the art learning based MVS models. On ETH3D, our method outperforms other recent learning-based approaches and performs comparably on advanced TnT.
****

Distinctiveness Oriented Positional Equilibrium for Point Cloud Registration
Taewon Min, Chonghyuk Song, Eunseok Kim, Inwook Shim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5490-5498
Recent state-of-the-art learning-based approaches to point cloud registration have largely been based on graph neural networks (GNN). However, these prominent GNN backbones suffer from the indistinguishable features problem associated with over-smoothing and structural ambiguity of the high-level features, a crucial bottleneck to point cloud registration that has evaded scrutiny in the recent relevant literature. To address this issue, we propose the Distinctiveness oriented Positional Equilibrium (DoPE) module, a novel positional embedding scheme that significantly improves the distinctiveness of the high-level features within both the source and target point clouds, resulting in superior point matching and hence registration accuracy. Specifically, we use the DoPE module in an iterative registration framework, whereby the two point clouds are gradually registered vi

a rigid transformations that are computed from DoPE's position-aware features. With every successive iteration, the DoPE module feeds increasingly consistent positional information to would-be corresponding pairs, which in turn enhances the resulting point-to-point correspondence predictions used to estimate the rigid transformation. Within only a few iterations, the network converges to a desired equilibrium, where the positional embeddings given to matching pairs become essentially identical. We validate the effectiveness of DoPE through comprehensive experiments on various registration benchmarks, registration task settings, and prominent backbones, yielding unprecedented performance improvement across all combinations.

****

Deep Edge-Aware Interactive Colorization Against Color-Bleeding Effects
Eungyeup Kim, Sanghyeon Lee, Jeonghoon Park, Somi Choi, Choonghyun Seo, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14667-14676
Deep neural networks for automatic image colorization often suffer from the color-bleeding artifact, a problematic color spreading near the boundaries between adjacent objects. Such color-bleeding artifacts debase the reality of generated outputs, limiting the applicability of colorization models in practice. Although previous approaches have attempted to address this problem in an automatic manner, they tend to work only in limited cases where a high contrast of gray-scale values are given in an input image. Alternatively, leveraging user interactions would be a promising approach for solving this color-breeding artifacts. In this paper, we propose a novel edge-enhancing network for the regions of interest via simple user scribbles indicating where to enhance. In addition, our method requires a minimal amount of effort from users for their satisfactory enhancement. Experimental results demonstrate that our interactive edge-enhancing approach effectively improves the color-bleeding artifacts compared to the existing baselines across various datasets.

****

ELSD: Efficient Line Segment Detector and Descriptor
Haotian Zhang, Yicheng Luo, Fangbo Qin, Yijia He, Xiao Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2969-2978
We present the novel Efficient Line Segment Detector and Descriptor (ELSD) to simultaneously detect line segments and extract their descriptors in an image. Unlike the traditional pipelines that conduct detection and description separately, ELSD utilizes a shared feature extractor for both detection and description, to provide the essential line features to the higher-level tasks like SLAM and image matching in real time. First, we design a one-stage compact model, and propose to use the mid-point, angle and length as the minimal representation of line segment, which also guarantees the center-symmetry. The non-centerness suppression is proposed to filter out the fragmented line segments caused by lines' intersections. The fine offset prediction is designed to refine the mid-point localization. Second, the line descriptor branch is integrated with the detector branch, and the two branches are jointly trained in an end-to-end manner. In the experiments, the proposed ELSD achieves the state-of-the-art performance on the Wirefrane dataset and YorkUrban dataset, in both accuracy and efficiency. The line description ability of ELSD also outperforms the previous works on the line matching task.

****

Separable Flow: Learning Motion Cost Volumes for Optical Flow Estimation
Feihu Zhang, Oliver J. Woodford, Victor Adrian Prisacariu, Philip H.S. Torr; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10807-10817
Full-motion cost volumes play a central role in current state-of-the-art optical flow methods. However, constructed using simple feature correlations, they lack the ability to encapsulate prior, or even non-local, knowledge. This creates artifacts in poorly constrained, ambiguous regions, such as occluded and textureless areas. We propose a separable cost volume module, a drop-in replacement to correlation cost volumes, that uses non-local aggregation layers to exploit global

context cues and prior knowledge, in order to disambiguate motions in these regions. Our method leads both the now standard Sintel and KITTI optical flow benchmarks in terms of accuracy, and is also shown to generalize better from synthetic to real data.

****

Learned Spatial Representations for Few-Shot Talking-Head Synthesis

Moustafa Meshry, Saksham Suri, Larry S. Davis, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13829-13838

We propose a novel approach for few-shot talking-head synthesis. While recent works in neural talking heads have produced promising results, they can still produce images that do not preserve the identity of the subject in source images. We posit this is a result of the entangled representation of each subject in a single latent code that models 3D shape information, identity cues, colors, lighting and even background details. In contrast, we propose to factorize the representation of a subject into its spatial and style components. Our method generates a target frame in two steps. First, it predicts a dense spatial layout for the target image. Second, an image generator utilizes the predicted layout for spatial denormalization and synthesizes the target frame. We experimentally show that this disentangled representation leads to a significant improvement over previous methods, both quantitatively and qualitatively.

****

Vision Transformers for Dense Prediction

René Ranftl, Alexey Bochkovskiy, Vladlen Koltun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12179-12188

We introduce dense prediction transformers, an architecture that leverages vision transformers in place of convolutional networks as a backbone for dense prediction tasks. We assemble tokens from various stages of the vision transformer into image-like representations at various resolutions and progressively combine them into full resolution predictions using a convolutional decoder. The transformer backbone processes representations at a constant and relatively high resolution and has a global receptive field at every stage. These properties allow the dense prediction transformer to provide finer-grained and more globally coherent predictions when compared to fully-convolutional networks. Our experiments show that this architecture yields substantial improvements on dense prediction tasks, especially when a large amount of training data is available. For monocular depth estimation, we observe an improvement of up to 28% in relative performance when compared to a state-of-the-art fully-convolutional network. When applied to semantic segmentation, dense prediction transformers set a new state of the art on ADE20K with 49.02% mIoU. We further show that the architecture can be fine-tuned on smaller datasets such as NYUv2, KITTI, and Pascal Context where it also sets the new state of the art. Our models are available at https://github.com/intel-isl/DPT.

****

V-DESIRR: Very Fast Deep Embedded Single Image Reflection Removal

B H Pawan Prasad, Green Rosh K S, Lokesh R. Boregowda, Kaushik Mitra, Sanjoy Chowdhury; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2390-2399

Real world images often gets corrupted due to unwanted reflections and their removal is highly desirable. A major share of such images originate from smart phone cameras capable of very high resolution captures. Most of the existing methods either focus on restoration quality by compromising on processing speed and memory requirements or, focus on removing reflections at very low resolutions, there by limiting their practical deploy-ability. We propose a light weight deep learning model for reflection removal using a novel scale space architecture. Our method processes the corrupted image in two stages, a Low Scale Sub-network (LSSNet) to process the lowest scale and a Progressive Inference (PI) stage to process all the higher scales. In order to reduce the computational complexity, the sub-networks in PI stage are designed to be much shallower than LSSNet. Moreover, we employ weight sharing between various scales within the PI stage to limit the

model size. This also allows our method to generalize to very high resolutions without explicit retraining. Our method is superior both qualitatively and quantitatively compared to the state of the art methods and at the same time 20x faster with 50x less number of parameters compared to the most recent state-of-the-art algorithm RAGNet. We implemented our method on an android smart phone, where a high resolution 12 MP image is restored in under 5 seconds.
****

CrackFormer: Transformer Network for Fine-Grained Crack Detection
Huajun Liu, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, Hui Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3783-3792
Cracks are irregular line structures that are of interest in many computer vision applications. Crack detection (e.g., from pavement images) is a challenging task due to intensity in-homogeneity, topology complexity, low contrast and noisy background. The overall crack detection accuracy can be significantly affected by the detection performance on fine-grained cracks. In this work, we propose a Crack Transformer network (CrackFormer) for fine-grained crack detection. The CrackFormer is composed of novel attention modules in a SegNet-like encoder-decoder architecture. Specifically, it consists of novel self-attention modules with 1x1 convolutional kernels for efficient contextual information extraction across feature-channels, and efficient positional embedding to capture large receptive field contextual information for long range interactions. It also introduces new scaling-attention modules to combine outputs from the corresponding encoder and decoder blocks to suppress non-semantic features and sharpen semantic cracks. The CrackFormer is trained and evaluated on three classical crack datasets. The experimental results show that CrackFormer achieves ODS values of 0.871, 0.877 and 0.881, respectively, on the three datasets and outperforms the state-of-the-art methods.
****

Factorizing Perception and Policy for Interactive Instruction Following
Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1888-1897
Performing simple household tasks based on language directives is very natural to humans, yet it remains an open challenge for an AI agent. The 'interactive instruction following' task attempts to make progress towards building an agent that can jointly navigate, interact, and reason in the environment at every step. To address the multifaceted problem, we propose a model that factorizes the task into interactive perception and action policy streams with enhanced components. We empirically validate that our model outperforms prior arts by significant margins on the ALFRED benchmark in all metrics with improved generalization.
****

Detecting Invisible People
Tarasha Khurana, Achal Dave, Deva Ramanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3174-3184
Monocular object detection and tracking have improved drastically in recent years, but rely on a key assumption: that objects are visible to the camera. Many offline tracking approaches reason about occluded objects post-hoc, by linking together tracklets after the object re-appears, making use of reidentification (ReID). However, online tracking in embodied robotic agents (such as a self-driving vehicle) fundamentally requires object permanence, which is the ability to reason about occluded objects before they re-appear. In this work, we re-purpose tracking benchmarks and propose new metrics for the task of detecting invisible objects, focusing on the illustrative case of people. We demonstrate that current detection and tracking systems perform dramatically worse on this task. We introduce two key innovations to recover much of this performance drop. We treat occluded object detection in temporal sequences as a short-term forecasting challenge, bringing to bear tools from dynamic sequence prediction. Second, we build dynamic models that explicitly reason in 3D from monocular videos without calibration, using observations produced by monocular depth estimators. To our knowledge, o

urs is the first work to demonstrate the effectiveness of monocular depth estima tion for the task of tracking and detecting occluded objects. Our approach stron gly improves by 11.4% over the baseline in ablations and by 5.0% over the state-of-the-art in F1 score.
****

GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds
Zekun Hao, Arun Mallya, Serge Belongie, Ming-Yu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14072-14082
We present GANcraft, an unsupervised neural rendering framework for generating p hotorealistic images of large 3D block worlds such as those created in Minecraft . Our method takes a semantic block world as input, where each block is assigned a semantic label such as dirt, grass, or water. We represent the world as a con tinuous volumetric function and train our model to render view-consistent photor ealistic images for a user-controlled camera. In the absence of paired ground tr uth real images for the block world, we devise a training technique based on pse udo-ground truth and adversarial training. This stands in contrast to prior work on neural rendering for view synthesis, which requires ground truth images to e stimate scene geometry and view-dependent appearance. In addition to camera traj ectory, GANcraft allows user control over both scene semantics and output style. Experimental results with comparison to strong baselines show the effectiveness of GANcraft on this novel task of photorealistic 3D block world synthesis.
****

Talk-To-Edit: Fine-Grained Facial Editing via Dialog
Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, Ziwei Liu; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 137 99-13808
Facial editing is an important task in vision and graphics with numerous applica tions. However, existing works are incapable to deliver a continuous and fine-gr ained editing mode (e.g., editing a slightly smiling face to a big laughing one) with natural interactions with users. In this work, we propose Talk-to-Edit, an interactive facial editing framework that performs fine-grained attribute manip ulation through dialog between the user and the system. Our key insight is to mo del a continual ""semantic field"" in the GAN latent space. 1) Unlike previous w orks that regard the editing as traversing straight lines in the latent space, h ere the fine-grained editing is formulated as finding a curving trajectory that respects fine-grained attribute landscape on the semantic field. 2) The curvatur e at each step is location-specific and determined by the input image as well as the users' language requests. 3) To engage the users in a meaningful dialog, ou r system generates language feedback by considering both the user request and th e current state of the semantic field. We also contribute CelebA-Dialog, a visua l-language facial editing dataset to facilitate large-scale study. Specifically, each image has manually annotated fine-grained attribute annotations as well as template-based textual descriptions in natural language. Extensive quantitative and qualitative experiments demonstrate the superiority of our framework in ter ms of 1) the smoothness of fine-grained editing, 2) the identity/attribute prese rvation, and 3) the visual photorealism and dialog fluency. Notably, user study validates that our overall system is consistently favored by around 80% of the p articipants.
****

AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting
Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris M. Kitani; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2021, pp. 9813-9823
Predicting accurate future trajectories of multiple agents is essential for auto nomous systems but is challenging due to the complex interaction between agents and the uncertainty in each agent's future behavior. Forecasting multi-agent tra jectories requires modeling two key dimensions: (1) time dimension, where we mod el the influence of past agent states over future states; (2) social dimension, where we model how the state of each agent affects others. Most prior methods mo del these two dimensions separately, e.g., first using a temporal model to summa rize features over time for each agent independently and then modeling the inter

action of the summarized features with a social model. This approach is suboptim al since independent feature encoding over either the time or social dimension c an result in a loss of information. Instead, we would prefer a method that allow s an agent's state at one time to directly affect another agent's state at a fut ure time. To this end, we propose a new Transformer, termed AgentFormer, that si multaneously models the time and social dimensions. The model leverages a sequen ce representation of multi-agent trajectories by flattening trajectory features across time and agents. Since standard attention operations disregard the agent identity of each element in the sequence, AgentFormer uses a novel agent-aware a ttention mechanism that preserves agent identities by attending to elements of t he same agent differently than elements of other agents. Based on AgentFormer, w e propose a stochastic multi-agent trajectory prediction model that can attend t o features of any agent at any previous timestep when inferring an agent's futur e position. The latent intent of all agents is also jointly modeled, allowing th e stochasticity in one agent's behavior to affect other agents. Extensive experi ments show that our method significantly improves the state of the art on well-e stablished pedestrian and autonomous driving datasets.

****

Transparent Object Tracking Benchmark

Heng Fan, Halady Akhilesha Miththanthaya, Harshit, Siranjiv Ramana Rajan, Xiaoqi ong Liu, Zhilin Zou, Yuewei Lin, Haibin Ling; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2021, pp. 10734-10743

Visual tracking has achieved considerable progress in recent years. However, cur rent research in the field mainly focuses on tracking of opaque objects, while l ittle attention is paid to transparent object tracking. In this paper, we make t he first attempt in exploring this problem by proposing a Transparent Object Tra cking Benchmark (TOTB). Specifically, TOTB consists of 225 videos (86K frames) f rom 15 diverse transparent object categories. Each sequence is manually labeled with axis-aligned bounding boxes. To the best of our knowledge, TOTB is the firs t benchmark dedicated to transparent object tracking. In order to understand how existing trackers perform and to provide comparison for future research on TOTB , we extensively evaluate 25 state-of-the-art tracking algorithms. The evaluatio n results exhibit that more efforts are needed to improve transparent object tra cking. Besides, we observe some nontrivial findings from the evaluation that are discrepant with some common beliefs in opaque object tracking. For example, we find that deep(er) features are not always good for improvements. Moreover, to e ncourage future research, we introduce a novel tracker, named TransATOM, which l everages transparency features for tracking and surpasses all 25 evaluated appro aches by a large margin. By releasing TOTB, we expect to facilitate future resea rch and application of transparent object tracking in both the academia and indu stry. The TOTB and evaluation results as well as TransATOM are available at http s://hengfan2010.github.io/projects/TOTB/.

****

Boosting Weakly Supervised Object Detection via Learning Bounding Box Adjusters

Bowen Dong, Zitong Huang, Yuelin Guo, Qilong Wang, Zhenxing Niu, Wangmeng Zuo; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 021, pp. 2876-2885

Weakly-supervised object detection (WSOD) has emerged as an inspiring recent top ic to avoid expensive instance-level object annotations. However, the bounding b oxes of most existing WSOD methods are mainly determined by precomputed proposal s, thereby being limited in precise object localization. In this paper, we defen d the problem setting for improving localization performance by leveraging the b ounding box regression knowledge from a well-annotated auxiliary dataset. First, we use the well-annotated auxiliary dataset to explore a series of learnable bo unding box adjusters (LBBAs) in a multi-stage training manner, which is class-ag nostic. Then, only LBBAs and a weakly-annotated dataset with non-overlapped clas ses are used for training LBBA-boosted WSOD. As such, our LBBAs are practically more convenient and economical to implement while avoiding the leakage of the au xiliary well-annotated dataset. In particular, we formulate learning bounding bo x adjusters as a bi-level optimization problem and suggest an EM-like multi-stag

e training algorithm. Then, a multi-stage scheme is further presented for LBBA-b oosted WSOD. Additionally, a masking strategy is adopted to improve proposal cla ssification. Experimental results verify the effectiveness of our method. Our me thod performs favorably against state-of-the-art WSOD methods and knowledge tran sfer model with similar problem setting. Code is publicly available at https://g ithub.com/DongSky/lbba_boosted_wsod.
****

Group-Wise Inhibition Based Feature Regularization for Robust Classification
Haozhe Liu, Haoqian Wu, Weicheng Xie, Feng Liu, Linlin Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 478-486
The convolutional neural network (CNN) is vulnerable to degraded images with eve n very small variations (e.g. corrupted and adversarial samples). One of the pos sible reasons is that CNN pays more attention to the most discriminative regions , but ignores the auxiliary features when learning, leading to the lack of featu re diversity for final judgment. In our method, we propose to dynamically suppre ss significant activation values of CNN by group-wise inhibition, but not fixedl y or randomly handle them when training. The feature maps with different activat ion distribution are then processed separately to take the feature independence into account. CNN is finally guided to learn richer discriminative features hier archically for robust classification according to the proposed regularization. O ur method is comprehensively evaluated under multiple settings, including classi fication against corruptions, adversarial attacks and low data regime. Extensive experimental results show that the proposed method can achieve significant impr ovements in terms of both robustness and generalization performances, when compa red with the state-of-the-art methods. Code is available at https://github. com/ LinusWu/TENET_Training.
****

Incorporating Convolution Designs Into Visual Transformers
Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, Wei Wu; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 579 -588
Motivated by the success of Transformers in natural language processing (NLP) ta sks, there exist some attempts (e.g., ViT and DeiT) to apply Transformers to the vision domain. However, pure Transformer architectures often require a large am ount of training data or extra supervision to obtain comparable performance with convolutional neural networks (CNNs). To overcome these limitations, we analyze the potential drawbacks when directly borrowing Transformer architectures from NLP. Then we propose a new Convolution-enhanced image Transformer (CeiT) which c ombines the advantages of CNNs in extracting low-level features, strengthening l ocality, and the advantages of Transformers in establishing long-range dependenc ies. Three modifications are made to the original Transformer: 1) instead of the straightforward tokenization from raw input images, we design an Image-to-Token s (I2T) module that extracts patches from generated low-level features; 2) the f eed-froward network in each encoder block is replaced with a Locally-enhanced Fe ed-Forward (LeFF) layer that promotes the correlation among neighboring tokens i n the spatial dimension; 3) a Layer-wise Class token Attention (LCA) is attached at the top of the Transformer that utilizes the multi-level representations. Ex perimental results on ImageNet and seven downstream tasks show the effectiveness and generalization ability compared with previous Transformers and state-of-the -art CNNs, without requiring a large amount of training data and extra CNN teach ers. Besides, CeiT models also demonstrate better convergence with 3xfewer train ing iterations, which can reduce the training cost significantly.
****

CDS: Cross-Domain Self-Supervised Pre-Training
Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, Kate Saenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9123-9132
We present a two-stage pre-training approach that improves the generalization ab ility of standard single-domain pre-training. While standard pre-training on a s ingle large dataset (such as ImageNet) can provide a good initial representation

for transfer learning tasks, this approach may result in biased representations that impact the success of learning with new multi-domain data (e.g., different artistic styles) via methods like domain adaptation. We propose a novel pre-training approach called Cross-Domain Self-supervision (CDS), which directly employs unlabeled multi-domain data for downstream domain transfer tasks. Our approach uses self-supervision not only within a single domain but also across domains. In-domain instance discrimination is used to learn discriminative features on new data in a domain-adaptive manner, while cross-domain matching is used to learn domain-invariant features. We apply our method as a second pre-training step (after ImageNet pre-training), resulting in a significant target accuracy boost to diverse domain transfer tasks compared to standard one-stage pre-training.

****

CaT: Weakly Supervised Object Detection With Category Transfer

Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, Yan-Feng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3070-3079

A large gap exists between fully-supervised object detection and weakly-supervised object detection. To narrow this gap, some methods consider knowledge transfer from additional fully-supervised dataset. But these methods do not fully exploit discriminative category information in the fully-supervised dataset, thus causing low mAP. To solve this issue, we propose a novel category transfer framework for weakly supervised object detection. The intuition is to fully leverage both visually-discriminative and semantically-correlated category information in the fully-supervised dataset to enhance the object-classification ability of a weakly-supervised detector. To handle overlapping category transfer, we propose a double-supervision mean teacher to gather common category information and bridge the domain gap between two datasets. To handle non-overlapping category transfer, we propose a semantic graph convolutional network to promote the aggregation of semantic features between correlated categories. Experiments are conducted with Pascal VOC 2007 as the target weakly-supervised dataset and COCO as the source fully-supervised dataset. Our category transfer framework achieves 63.5% mAP and 80.3% CorLoc with 5 overlapping categories between two datasets, which outperforms the state-of-the-art methods. Codes are avaliable at https://github.com/MediaBrain-SJTU/CaT.

****

4DComplete: Non-Rigid Motion Estimation Beyond the Observable Surface

Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12706-12716

Tracking non-rigidly deforming scenes using range sensors has numerous applications including computer vision, AR/VR, and robotics. However, due to occlusions and physical limitations of range sensors, existing methods only handle the visible surface, thus causing discontinuities and incompleteness in the motion field. To this end, we introduce 4DComplete, a novel data-driven approach that estimates the non-rigid motion for the unobserved geometry. 4DComplete takes as input a partial shape and motion observation, extracts 4D time-space embedding, and jointly infers the missing geometry and motion field using a sparse fully-convolutional network. For network training, we constructed a large-scale synthetic dataset called DeformingThings4D, which consists of 1,972 animation sequences spanning 31 different animals or humanoid categories with dense 4D annotation. Experiments show that 4DComplete 1) reconstructs high-resolution volumetric shape and motion field from a partial observation, 2) learns an entangled 4D feature representation that benefits both shape and motion estimation, 3) yields more accurate and natural deformation than classic non-rigid priors such as As-RigidAs-Possible (ARAP) deformation, and 4) generalizes well to unseen objects in real-world sequences.

****

Scaling Semantic Segmentation Beyond 1K Classes on a Single GPU

Shipra Jain, Danda Pani Paudel, Martin Danelljan, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7426-7

The state-of-the-art object detection and image classification methods can perfo rm impressively on more than 9k and 10k classes respectively. In contrast, the n umber of classes in semantic segmentation datasets is relatively limited. This i s not surprising when the restrictions caused by the lack of labelled data and h igh computation demand for segmentation are considered. In this paper, we propos e a novel training methodology to train and scale the existing semantic segmenta tion models for a large number of semantic classes without increasing the memory overhead. In our approach, we reduce the space complexity of the segmentation m odel's output from O(C) to O(1), propose an approximation method for ground-trut h class probability, and use it to compute cross-entropy loss. The proposed appr oach is general and can be adopted by any state-of-the-art segmentation model to gracefully scale it for any number of semantic classes with only one GPU. Our a pproach achieves similar, and in some cases even better mIoU for Cityscapes, Pas cal VOC and ADE20k dataset when adopted to DeeplabV3+ model with different backb ones. We demonstrate a clear benefit of our approach on a dataset with 1284 clas ses, bootstrapped from LVIS and COCO annotations, with almost three times better mIoU when compared to DeeplabV3+. Code is available at: https://github.com/ship ra25jain/ESSNet.

****

Searching for Robustness: Loss Learning for Noisy Classification Tasks
Boyan Gao, Henry Gouk, Timothy M. Hospedales; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2021, pp. 6670-6679
We present a "learning to learn" approach for discovering white-box classificati on loss functions that are robust to label noise in the training data. We parame terise a flexible family of loss functions using Taylor polynomials, and apply e volutionary strategies to search for noise-robust losses in this space. To learn re-usable loss functions that can apply to new tasks, our fitness function scor es their performance in aggregate across a range of training datasets and archit ectures. The resulting white-box loss provides a simple and fast "plug-and-play" module that enables effective label-noise-robust learning in diverse downstream tasks, without requiring a special training procedure or network architecture. The efficacy of our loss is demonstrated on a variety of datasets with both synt hetic and real label noise, where we compare favourably to prior work.

****

On Compositions of Transformations in Contrastive Self-Supervised Learning
Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 9577-9587
In the image domain, excellent representations can be learned by inducing invari ance to content-preserving transformations via noise contrastive learning. In th is paper, we generalize contrastive learning to a wider set of transformations, and their compositions, for which either invariance or distinctiveness is sought . We show that it is not immediately obvious how existing methods such as SimCLR can be extended to do so. Instead, we introduce a number of formal requirements that all contrastive formulations must satisfy, and propose a practical constru ction which satisfies these requirements. In order to maximise the reach of this analysis, we express all components of noise contrastive formulations as the ch oice of certain generalized transformations of the data (GDTs), including data s ampling. We then consider videos as an example of data in which a large variety of transformations are applicable, accounting for the extra modalities -- for wh ich we analyze audio and text -- and the dimension of time. We find that being i nvariant to certain transformations and distinctive to others is critical to lea rning effective video representations, improving the state-of-the-art for multip le benchmarks by a large margin, and even surpassing supervised pretraining.

****

Handwriting Transformers
Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Sha hbaz Khan, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1086-1094

We propose a novel transformer-based styled handwritten text image generation ap
proach, HWT, that strives to learn both style-content entanglement as well as gl
obal and local style patterns. The proposed HWT captures the long and short rang
e relationships within the style examples through a self-attention mechanism, th
ereby encoding both global and local style patterns. Further, the proposed trans
former-based HWT comprises an encoder-decoder attention that enables style-conte
nt entanglement by gathering the style features of each query character. To the
best of our knowledge, we are the first to introduce a transformer-based network
 for styled handwritten text generation. Our proposed HWT generates realistic st
yled handwritten text images and outperforms the state-of-the-art demonstrated t
hrough extensive qualitative, quantitative and human-based evaluations. The prop
osed HWT can handle arbitrary length of text and any desired writing style in a
few-shot setting. Further, our HWT generalizes well to the challenging scenario
where both words and writing style are unseen during training, generating realis
tic styled handwritten text images.
****

BV-Person: A Large-Scale Dataset for Bird-View Person Re-Identification
Cheng Yan, Guansong Pang, Lei Wang, Jile Jiao, Xuetao Feng, Chunhua Shen, Jingji
ng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2021, pp. 10943-10952
Person Re-IDentification (ReID) aims at re-identifying persons from non-overlapp
ing cameras. Existing person ReID studies focus on horizontal-view ReID tasks, i
n which the person images are captured by the cameras from a (nearly) horizontal
 view. In this work we introduce a new ReID task, bird-view person ReID, which a
ims at searching for a person in a gallery of horizontal-view images with the qu
ery images taken from a bird's-eye view, i.e., an elevated view of an object fro
m above. The task is important because there are a large number of video surveil
lance cameras capturing persons from such an elevated view at public places. How
ever, it is a challenging task in that the images from the bird view (i) provide
 limited person appearance information and (ii) have a large discrepancy compare
d to the persons in the horizontal view. We aim to facilitate the development of
 person ReID from this line by introducing a large-scale real-world dataset for
this task. The proposed dataset, named BV-Person, contains 114k images of 18k id
entities in which nearly 20k images of 7.4k identities are taken from the bird's
-eye view. We further introduce a novel model for this new ReID task. Large-scal
e experiments are performed to evaluate our model and 11 current state-of-the-ar
t ReID models on BV-Person to establish performance benchmarks from multiple per
spectives. The empirical results show that our model consistently and substantia
lly outperforms the state-of-the-arts on all five datasets derived from BV-Perso
n. Our model also achieves state-of-the-art performance on two general ReID data
sets. Our code and dataset will be made publicly available.
****

Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Est
imation From Images in the Wild
Akash Sengupta, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 11219-11229
This paper addresses the problem of 3D human body shape and pose estimation from
 an RGB image. This is often an ill-posed problem, since multiple plausible 3D b
odies may match the visual evidence present in the input - particularly when the
 subject is occluded. Thus, it is desirable to estimate a distribution over 3D b
ody shape and pose conditioned on the input image instead of a single 3D reconst
ruction. We train a deep neural network to estimate a hierarchical matrix-Fisher
 distribution over relative 3D joint rotation matrices (i.e. body pose), which e
xploits the human body's kinematic tree structure, as well as a Gaussian distrib
ution over SMPL body shape parameters. To further ensure that the predicted shap
e and pose distributions match the visual evidence in the input image, we implem
ent a differentiable rejection sampler to impose a reprojection loss between gro
und-truth 2D joint coordinates and samples from the predicted distributions, pro
jected onto the image plane. We show that our method is competitive with the sta
te-of-the-art in terms of 3D shape and pose metrics on the SSP-3D and 3DPW datas

ets, while also yielding a structured probability distribution over 3D body shape and pose, with which we can meaningfully quantify prediction uncertainty and sample multiple plausible 3D reconstructions to explain a given input image.
****

Dynamic DETR: End-to-End Object Detection With Dynamic Attention
Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2988-2997
In this paper, we present a novel Dynamic DETR (Detection with Transformers) approach by introducing dynamic attentions into both the encoder and decoder stages of DETR to break its two limitations on small feature resolution and slow training convergence. To address the first limitation, which is due to the quadratic computational complexity of the self-attention module in Transformer encoders, we propose a dynamic encoder to approximate the Transformer encoder's attention mechanism using a convolution-based dynamic encoder with various attention types. Such an encoder can dynamically adjust attentions based on multiple factors such as scale importance, spatial importance, and representation (i.e., feature dimension) importance. To mitigate the second limitation of learning difficulty, we introduce a dynamic decoder by replacing the cross-attention module with a ROI-based dynamic attention in the Transformer decoder. Such a decoder effectively assists Transformers to focus on region of interests from a coarse-to-fine manner and dramatically lowers the learning difficulty, leading to a much faster convergence with fewer training epochs. We conduct a series of experiments to demonstrate our advantages. Our Dynamic DETR significantly reduces the training epochs (by \bf 14x), yet results in a much better performance (by \bf 3.6 on mAP). Meanwhile, in the standard 1x setup with ResNet-50 backbone, we archive a new state-of-the-art performance that further proves the learning effectiveness of the proposed approach. Code will be released soon.
****

DepthTrack: Unveiling the Power of RGBD Tracking
Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, Joni-Kristian Kämäräinen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10725-10733
RGBD (RGB plus depth) object tracking is gaining momentum as RGBD sensors have become popular in many application fields such as robotics. However, the best RGBD trackers are extensions of the state-of-the-art deep RGB trackers. They are trained with RGB data and the depth channel is used as a sidekick for subtleties such as occlusion detection. This can be explained by the fact that there are no sufficiently large RGBD datasets to 1) train "deep depth trackers" and to 2) challenge RGB trackers with sequences for which the depth cue is essential. This work introduces a new RGBD tracking dataset - DepthTrack - that has twice as many sequences (200) and scene types (40) than in the largest existing dataset, and three times more objects (90). In addition, the average length of the sequences (1473), the number of deformable objects (16) and the number of annotated tracking attributes (15) have been increased. Furthermore, by running the SotA RGB and RGBD trackers on DepthTrack, we propose a new RGBD tracking baseline, namely DeT, which reveals that deep RGBD tracking indeed benefits from genuine training data. The code and dataset is available at https://github.com/xiaozai/DeT.
****

XVFI: eXtreme Video Frame Interpolation
Hyeonjun Sim, Jihyong Oh, Munchurl Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14489-14498
In this paper, we firstly present a dataset (X4K1000FPS) of 4K videos of 1000 fps with the extreme motion to the research community for video frame interpolation (VFI), and propose an extreme VFI network, called XVFI-Net, that first handles the VFI for 4K videos with large motion. The XVFI-Net is based on a recursive multi-scale shared structure that consists of two cascaded modules for bidirectional optical flow learning between two input frames (BiOF-I) and for bidirectional optical flow learning from target to input frames (BiOF-T). The optical flows are stably approximated by a complementary flow reversal (CFR) proposed in BiOF-

T module. During inference, the BiOF-I module can start at any scale of input wh
ile the BiOF-T module only operates at the original input scale so that the infe
rence can be accelerated while maintaining highly accurate VFI performance. Exte
nsive experimental results show that our XVFI-Net can successfully capture the e
ssential information of objects with extremely large motions and complex texture
s while the state-of-the-art methods exhibit poor performance. Furthermore, our
XVFI-Net framework also performs comparably on the previous lower resolution ben
chmark dataset, which shows a robustness of our algorithm as well. All source co
des, pre-trained models, and proposed X4K1000FPS datasets are publicly available
at https://github.com/JihyongOh/XVFI.
****

Cortical Surface Shape Analysis Based on Alexandrov Polyhedra
Min Zhang, Yang Guo, Na Lei, Zhou Zhao, Jianfeng Wu, Xiaoyin Xu, Yalin Wang, Xia
nfeng Gu; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2021, pp. 14244-14252
Shape analysis has been playing an important role in early diagnosis and prognos
is of neurodegenerative diseases such as Alzheimer's diseases (AD). However, obt
aining effective shape representations remains challenging. This paper proposes
to use the Alexandrov polyhedra as surface-based shape signatures for cortical m
orphometry analysis. Given a closed genus-0 surface, its Alexandrov polyhedron i
s a convex representation that encodes its intrinsic geometry information. We pr
opose to compute the polyhedra via a novel spherical optimal transport (OT) comp
utation. In our experiments, we observe that the Alexandrov polyhedra of cortica
l surfaces between pathology-confirmed AD and cognitively unimpaired individuals
are significantly different. Moreover, we propose a visualization method by com
paring local geometry differences across cortical surfaces. We show that the pro
posed method is effective in pinpointing regional cortical structural changes im
pacted by AD.
****

Watch Only Once: An End-to-End Video Action Detection Framework
Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen,
Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2021, pp. 8178-8187
We propose an end-to-end pipeline, named Watch Once Only (WOO), for video action
detection. Current methods either decouple video action detection task into sep
arated stages of actor localization and action classification or train two separ
ated models within one stage. In contrast, our approach solves the actor localiz
ation and action classification simultaneously in a unified network. The whole p
ipeline is significantly simplified by unifying the backbone network and elimina
ting many hand-crafted components. WOO takes a unified video backbone to simulta
neously extract features for actor location and action classification. In additi
on, we introduce spatial-temporal action embeddings into our framework and desig
n a spatial-temporal fusion module to obtain more discriminative features with r
icher information, which further boosts the action classification performance. E
xtensive experiments on AVA and JHMDB datasets show that WOO achieves state-of-t
he-art performance, while still reduces up to 16.7% GFLOPs compared with existin
g methods. We hope our work can inspire rethinking the convention of action dete
ction and serve as a solid baseline for end-to-end action detection. Code is ava
ilable.
****

Spatial-Temporal Consistency Network for Low-Latency Trajectory Forecasting
Shijie Li, Yanying Zhou, Jinhui Yi, Juergen Gall; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2021, pp. 1940-1949
Trajectory forecasting is a crucial step for autonomous vehicles and mobile robo
ts in order to navigate and interact safely. In order to handle the spatial inte
ractions between objects, graph-based approaches have been proposed. These metho
ds, however, model motion on a frame-to-frame basis and do not provide a strong
temporal model. To overcome this limitation, we propose a compact model called S
patial-Temporal Consistency Network (STC-Net). In STC-Net, dilated temporal conv
olutions are introduced to model long-range dependencies along each trajectory f

or better temporal modeling while graph convolutions are employed to model the s patial interaction among different trajectories. Furthermore, we propose a featu re-wise convolution to generate the predicted trajectories in one pass and refin e the forecast trajectories together with the reconstructed observed trajectorie s. We demonstrate that STC-Net generates spatially and temporally consistent tra jectories and outperforms other graph-based methods. Since STC-Net requires only 0.7k parameters and forecasts the future with a latency of only 1.3ms, it advan ces the state-of-the-art and satisfies the requirements for realistic applicatio ns.
****

Point Cloud Augmentation With Weighted Local Transformations
Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, Hyunwoo J. Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2021, pp. 548-557
Despite the extensive usage of point clouds in 3D vision, relatively limited dat a are available for training deep neural networks. Although data augmentation is a standard approach to compensate for the scarcity of data, it has been less ex plored in the point cloud literature. In this paper, we propose a simple and eff ective augmentation method called PointWOLF for point cloud augmentation. The pr oposed method produces smoothly varying non-rigid deformations by locally weight ed transformations centered at multiple anchor points. The smooth deformations a llow diverse and realistic augmentations. Furthermore, in order to minimize the manual efforts to search the optimal hyperparameters for augmentation, we presen t AugTune, which generates augmented samples of desired difficulties producing t argeted confidence scores. Our experiments show that our framework consistently improves the performance for both shape classification and part segmentation tas ks. In particular, with PointNet++, PointWOLF achieves the state-of-the-art 89.7 accuracy on shape classification with the real-world ScanObjectNN dataset. The code is available at https://github.com/mlvlab/PointWOLF.
****

Solving Inefficiency of Self-Supervised Representation Learning
Guangrun Wang, Keze Wang, Guangcong Wang, Philip H.S. Torr, Liang Lin; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9505-9515
Self-supervised learning (especially contrastive learning) has attracted great i nterest due to its huge potential in learning discriminative representations in an unsupervised manner. Despite the acknowledged successes, existing contrastive learning methods suffer from very low learning efficiency, e.g., taking about t en times more training epochs than supervised learning for comparable recognitio n accuracy. In this paper, we reveal two contradictory phenomena in contrastive learning that we call under-clustering and over-clustering problems, which are m ajor obstacles to learning efficiency. Under-clustering means that the model can not efficiently learn to discover the dissimilarity between inter-class samples when the negative sample pairs for contrastive learning are insufficient to diff erentiate all the actual object classes. Over-clustering implies that the model cannot efficiently learn features from excessive negative sample pairs, forcing the model to over-cluster samples of the same actual classes into different clus ters. To simultaneously overcome these two problems, we propose a novel self-sup ervised learning framework using a truncated triplet loss. Precisely, we employ a triplet loss tending to maximize the relative distance between the positive pa ir and negative pairs to address the under-clustering problem; and we construct the negative pair by selecting a negative sample deputy from all negative sample s to avoid the over-clustering problem, guaranteed by the Bernoulli Distribution model. We extensively evaluate our framework in several large-scale benchmarks (e.g., ImageNet, SYSU-30k, and COCO). The results demonstrate our model's superi ority (e.g., the learning efficiency) over the latest state-of-the-art methods b y a clear margin. See Codes at: https://github.com/wanggrun/triplet.
****

Stochastic Scene-Aware Motion Prediction
Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, Mi

chael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11374-11384
A long-standing goal in computer vision is to capture, model, and realistically synthesize human behavior. Specifically, by learning from data, our goal is to enable virtual humans to navigate within cluttered indoor scenes and naturally interact with objects. Such embodied behavior has applications in virtual reality, computer games, and robotics, while synthesized behavior can be used as a source of training data. This is challenging because real human motion is diverse and adapts to the scene. For example, a person can sit or lie on a sofa in many places and with varying styles. It is necessary to model this diversity when synthesizing virtual humans that realistically perform human-scene interactions. We present a novel data-driven, stochastic motion synthesis method that models different styles of performing a given action with a target object. Our method, called SAMP, for Scene-Aware Motion Prediction, generalizes to target objects of various geometries while enabling the character to navigate in cluttered scenes. To train our method, we collected MoCap data covering various sitting, lying down, walking, and running styles. We demonstrate our method on complex indoor scenes and achieve superior performance compared to existing solutions. Our code and data are available for research at https://samp.is.tue.mpg.de.
****
Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation
Gwangbin Bae, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13137-13146
Surface normal estimation from a single image is an important task in 3D scene understanding. In this paper, we address two limitations shared by the existing methods: the inability to estimate the aleatoric uncertainty and lack of detail in the prediction. The proposed network estimates the per-pixel surface normal probability distribution. We introduce a new parameterization for the distribution, such that its negative log-likelihood is the angular loss with learned attenuation. The expected value of the angular error is then used as a measure of the aleatoric uncertainty. We also present a novel decoder framework where pixel-wise multi-layer perceptrons are trained on a subset of pixels sampled based on the estimated uncertainty. The proposed uncertainty-guided sampling prevents the bias in training towards large planar surfaces and improves the quality of prediction, especially near object boundaries and on small structures. Experimental results show that the proposed method outperforms the state-of-the-art in ScanNet and NYUv2, and that the estimated uncertainty correlates well with the prediction error. Code is available at https://github.com/baegwangbin/surface_normal_uncertainty.
****
Explaining in Style: Training a GAN To Explain a Classifier in StyleSpace
Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, Inbar Mosseri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 693-702
Image classification models can depend on multiple different semantic attributes of the image. An explanation of the decision of the classifier needs to both discover and visualize these properties. Here we present StylEx, a method for doing this, by training a generative model to specifically explain multiple attributes that underlie classifier decisions. A natural source for such attributes is the StyleSpace of StyleGAN, which is known to generate semantically meaningful dimensions in the image. However, because standard GAN training is not dependent on the classifier, it may not represent those attributes which are important for the classifier decision, and the dimensions of StyleSpace may represent irrelevant attributes. To overcome this, we propose a training procedure for a StyleGAN, which incorporates the classifier model, in order to learn a classifier-specific StyleSpace. Explanatory attributes are then selected from this space. These can be used to visualize the effect of changing multiple attributes per image, thus providing image-specific explanations. We apply StylEx to multiple domains, including animals, leaves, faces and retinal images. For these, we show how an ima

ge can be modified in different ways to change its classifier output. Our results show that the method finds attributes that align well with semantic ones, generate meaningful image-specific explanations, and are human-interpretable as measured in user-studies.

****

Exploring Visual Engagement Signals for Representation Learning

Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, Ser-Nam Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4206-4217

Visual engagement in social media platforms comprises interactions with photo posts including comments, shares, and likes. In this paper, we leverage such visual engagement clues as supervisory signals for representation learning. However, learning from engagement signals is non-trivial as it is not clear how to bridge the gap between low-level visual information and high-level social interaction. We present VisE,, a weakly supervised learning approach, which maps social images to pseudo labels derived by clustered engagement signals. We then study how models trained in this way benefit subjective downstream computer vision tasks such as emotion recognition or political bias detection. Through extensive studies, we empirically demonstrate the effectiveness of VisE across a diverse set of classification tasks beyond the scope of conventional recognition.

****

MUSIQ: Multi-Scale Image Quality Transformer

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, Feng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5148-5157

Image quality assessment (IQA) is an important research topic for understanding and improving visual experience. The current state-of-the-art IQA methods are based on convolutional neural networks (CNNs). The performance of CNN-based models is often compromised by the fixed shape constraint in batch training. To accommodate this, the input images are usually resized and cropped to a fixed shape, causing image quality degradation. To address this, we design a multi-scale image quality Transformer (MUSIQ) to process native resolution images with varying sizes and aspect ratios. With a multi-scale image representation, our proposed method can capture image quality at different granularities. Furthermore, a novel hash-based 2D spatial embedding and a scale embedding is proposed to support the positional embedding in the multi-scale representation. Experimental results verify that our method can achieve state-of-the-art performance on multiple large scale IQA datasets such as PaQ-2-PiQ, SPAQ and KonIQ-10k.

****

FcaNet: Frequency Channel Attention Networks

Zequn Qin, Pengyi Zhang, Fei Wu, Xi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 783-792

Attention mechanism, especially channel attention, has gained great success in the computer vision field. Many works focus on how to design efficient channel attention mechanisms while ignoring a fundamental problem, i.e., channel attention mechanism uses scalar to represent channel, which is difficult due to massive information loss. In this work, we start from a different view and regard the channel representation problem as a compression process using frequency analysis. Based on the frequency analysis, we mathematically prove that the conventional global average pooling is a special case of the feature decomposition in the frequency domain. With the proof, we naturally generalize the compression of the channel attention mechanism in the frequency domain and propose our method with multi-spectral channel attention, termed as FcaNet. FcaNet is simple but effective. We can change a few lines of code in the calculation to implement our method within existing channel attention methods. Moreover, the proposed method achieves state-of-the-art results compared with other channel attention methods on image classification, object detection, and instance segmentation tasks. Our method could consistently outperform the baseline SENet, with the same number of parameters and the same computational cost. Our code and models are publicly available at https://github.com/cfzd/FcaNet.

****
Matching in the Dark: A Dataset for Matching Image Pairs of Low-Light Scenes
Wenzheng Song, Masanori Suganuma, Xing Liu, Noriyuki Shimobayashi, Daisuke Maruta, Takayuki Okatani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6029-6038
This paper considers matching images of low-light scenes, aiming to widen the frontier of SfM and visual SLAM applications. Recent image sensors can record the brightness of scenes with more than eight-bit precision, available in their RAW-format image. We are interested in making full use of such high-precision information to match extremely low-light scene images that conventional methods cannot handle. For extreme low-light scenes, even if some of their brightness information exists in the RAW format images' low bits, the standard raw image processing fails to utilize them properly. As was recently shown by Chen et al., CNNs can learn to produce images with a natural appearance from such RAW-format images. To consider if and how well we can utilize such information stored in RAW-format images for image matching, we have created a new dataset named MID (matching in the dark). Using it, we experimentally evaluated combinations of eight image-enhancing methods and eleven image matching methods consisting of classical/neural local descriptors and classical/neural initial point-matching methods. The results show the advantage of using the RAW-format images and the strengths and weaknesses of the above component methods. They also imply there is room for further research.
****
ReDAL: Region-Based and Diversity-Aware Active Learning for Point Cloud Semantic Segmentation
Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, Winston H. Hsu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15510-15519
Despite the success of deep learning on supervised point cloud semantic segmentation, obtaining large-scale point-by-point manual annotations is still a significant challenge. To reduce the huge annotation burden, we propose a Region-based and Diversity-aware Active Learning (ReDAL), a general framework for many deep learning approaches, aiming to automatically select only informative and diverse sub-scene regions for label acquisition. Observing that only a small portion of annotated regions are sufficient for 3D scene understanding with deep learning, we use softmax entropy, color discontinuity, and structural complexity to measure the information of sub-scene regions. A diversity-aware selection algorithm is also developed to avoid redundant annotations resulting from selecting informative but similar regions in a querying batch. Extensive experiments show that our method highly outperforms previous active learning strategies, and we achieve the performance of 90% fully supervised learning, while less than 15% and 5% annotations are required on S3DIS and SemanticKITTI datasets, respectively.
****
Point Transformer
Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, Vladlen Koltun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16259-16268
Self-attention networks have revolutionized natural language processing and are making impressive strides in image analysis tasks such as image classification and object detection. Inspired by this success, we investigate the application of self-attention networks to 3D point cloud processing. We design self-attention layers for point clouds and use these to construct self-attention networks for tasks such as semantic scene segmentation, object part segmentation, and object classification. Our Point Transformer design improves upon prior work across domains and tasks. For example, on the challenging S3DIS dataset for large-scale semantic scene segmentation, the Point Transformer attains an mIoU of 70.4% on Area 5, outperforming the strongest prior model by 3.3 absolute percentage points and crossing the 70% mIoU threshold for the first time.
****
Self-Motivated Communication Agent for Real-World Vision-Dialog Navigation

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, Jianbin Jiao ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2021, pp. 1594-1603

Vision-Dialog Navigation (VDN) requires an agent to ask questions and navigate f ollowing the human responses to find target objects. Conventional approaches are only allowed to ask questions at predefined locations, which are built upon exp ensive dialogue annotations, and inconvenience the real-word human-robot communi cation and cooperation. In this paper, we propose a Self-Motivated Communication Agent (SCoA) that learns whether and what to communicate with human adaptively to acquire instructive information for realizing dialogue annotation-free naviga tion and enhancing the transferability in real-world unseen environment. Specifi cally, we introduce a whether-to-ask (WeTA) policy, together with uncertainty of which action to choose, to indicate whether the agent should ask a question. Th en, a what-to-ask (WaTA) policy is proposed, in which, along with the oracle's a nswers, the agent learns to score question candidates so as to pick up the most informative one for navigation, and meanwhile mimic oracle's answering. Thus, th e agent can navigate in a self-Q&A manner even in real-world environment where t he human assistance is often unavailable. Through joint optimization of communic ation and navigation in a unified imitation learning and reinforcement learning framework, SCoA asks a question if necessary and obtains a hint for guiding the agent to move towards the target with less communication cost. Experiments on se en and unseen environments demonstrate that SCoA shows not only superior perform ance over existing baselines without dialog annotations, but also competing resu lts compared with rich dialog annotations based counterparts.

****

Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation

Shu Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, Xiaoxing Zhang; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1564-1573

How to make the appearance and motion information interact effectively to accomm odate complex scenarios is a fundamental issue in flow-based zero-shot video obj ect segmentation. In this paper, we propose an Attentive Multi-Modality Collabor ation Network (AMC-Net) to utilize appearance and motion information uniformly. Specifically, AMC-Net fuses robust information from multi-modality features and promotes their collaboration in two stages. First, we propose a Multi-Modality C o-Attention Gate (MCG) on the bilateral encoder branches, in which a gate functi on is used to formulate co-attention scores for balancing the contributions of m ulti-modality features and suppressing the redundant and misleading information. Then, we propose a Motion Correction Module (MCM) with a visual-motion attentio n mechanism, which is constructed to emphasize the features of foreground object s by incorporating the spatio-temporal correspondence between appearance and mot ion cues. Extensive experiments on three public challenging benchmark datasets v erify that our proposed network performs favorably against existing state-of-the -art methods via training with fewer data.

****

Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis

Ajay Jain, Matthew Tancik, Pieter Abbeel; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2021, pp. 5885-5894

We present DietNeRF, a 3D neural scene representation estimated from a few image s. Neural Radiance Fields (NeRF) learn a continuous volumetric representation of a scene through multi-view consistency, and can be rendered from novel viewpoin ts by ray casting. While NeRF has an impressive ability to reconstruct geometry and fine details given many images, up to 100 for challenging 360 degree scenes, it often finds a degenerate solution to its image reconstruction objective when only a few input views are available. To improve few-shot quality, we propose D ietNeRF. We introduce an auxiliary semantic consistency loss that encourages rea listic renderings at novel poses. DietNeRF is trained on individual scenes to (1) correctly render given input views from the same pose, and (2) match high-leve l semantic attributes across different, random poses. Our semantic loss allows u s to supervise DietNeRF from arbitrary poses. We extract these semantics using a

pre-trained visual encoder such as CLIP, a Vision Transformer trained on hundreds of millions of diverse single-view, 2D photographs mined from the web with natural language supervision. In experiments, DietNeRF improves the perceptual quality of few-shot view synthesis when learned from scratch, can render novel views with as few as one observed image when pre-trained on a multi-view dataset, and produces plausible completions of completely unobserved regions. Our project website is available at https://www.ajayj.com/dietnerf.
****

CrossDet: Crossline Representation for Object Detection

Heqian Qiu, Hongliang Li, Qingbo Wu, Jianhua Cui, Zichen Song, Lanxiao Wang, Minjian Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3195-3204

Object detection aims to accurately locate and classify objects in an image, which requires precise object representations. Existing methods usually use rectangular anchor boxes or a set of points to represent objects. However, these methods either introduce background noise or miss the continuous appearance information inside the object, and thus cause incorrect detection results. In this paper, we propose a novel anchor-free object detection network, called CrossDet, which uses a set of growing cross lines along horizontal and vertical axes as object representations. An object can be flexibly represented as cross lines in different combinations. It not only can effectively reduce the interference of noise, but also takes into account the continuous object information, which is useful to enhance the discriminability of object features and find the object boundaries. Based on the learned cross lines, we propose a crossline extraction module to adaptively capture features of cross lines. Furthermore, we design a decoupled regression mechanism to regress the localization along the horizontal and vertical directions respectively, which helps to decrease the optimization difficulty because the optimization space is limited to a specific direction. Our method achieves consistently improvement on the PASCAL VOC and MS-COCO datasets. The experiment results demonstrate the effectiveness of our proposed method.
****

Graph-to-3D: End-to-End Generation and Manipulation of 3D Scenes Using Scene Graphs

Helisa Dhamo, Fabian Manhardt, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16352-16361

Controllable scene synthesis consists of generating 3D information that satisfy underlying specifications. Thereby, these specifications should be abstract, i.e. allowing easy user interaction, whilst providing enough interface for detailed control. Scene graphs are representations of a scene, composed of objects (nodes) and inter-object relationships (edges), proven to be particularly suited for this task, as they allow for semantic control on the generated content. Previous works tackling this task often rely on synthetic data, and retrieve object meshes, which naturally limits the generation capabilities. To circumvent this issue, we instead propose the first work that directly generates shapes from a scene graph in an end-to-end manner. In addition, we show that the same model supports scene modification, using the respective scene graph as interface. Leveraging Graph Convolutional Networks (GCN) we train a variational Auto-Encoder on top of the object and edge categories, as well as 3D shapes and scene layouts, allowing latter sampling of new scenes and shapes.
****

Universal and Flexible Optical Aberration Correction Using Deep-Prior Based Deconvolution

Xiu Li, Jinli Suo, Weihang Zhang, Xin Yuan, Qionghai Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2613-2621

High quality imaging usually requires bulky and expensive lenses to compensate geometric and chromatic aberrations. This poses high constraints on the optical hash or low cost applications. Although one can utilize algorithmic reconstruction to remove the artifacts of low-end lenses, the degeneration from optical aberrations is spatially varying and the computation has to trade off efficiency for

performance. For example, we need to conduct patch-wise optimization or train a large set of local deep neural networks to achieve high reconstruction performance across the whole image. In this paper, we propose a PSF aware plug-and-play deep network, which takes the aberrant image and PSF map as input and produces the latent high quality version via incorporating lens-specific deep priors, thus leading to a universal and flexible optical aberration correction method. Specifically, we pre-train a base model from a set of diverse lenses and then adapt it to a given lens by quickly refining the parameters, which largely alleviates the time and memory consumption of model learning. The approach is of high efficiency in both training and testing stages. Extensive results verify the promising applications of our proposed approach for compact low-end cameras.
****

E-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks
Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, Thomas Lukasiewicz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1244-1254
Recently, there has been an increasing number of efforts to introduce models capable of generating natural language explanations (NLEs) for their predictions on vision-language (VL) tasks. Such models are appealing, because they can provide human-friendly and comprehensive explanations. However, there is a lack of comparison between existing methods, which is due to a lack of re-usable evaluation frameworks and a scarcity of datasets. In this work, we introduce e-ViL and e-SNLI-VE. e-ViL is a benchmark for explainable vision-language tasks that establishes a unified evaluation framework and provides the first comprehensive comparison of existing approaches that generate NLEs for VL tasks. It spans four models and three datasets and both automatic metrics and human evaluation are used to assess model-generated explanations. e-SNLI-VE is currently the largest existing VL dataset with NLEs (over 430k instances). We also propose a new model that combines UNITER, which learns joint embeddings of images and text, and GPT-2, a pre-trained language model that is well-suited for text generation. It surpasses the previous state of the art by a large margin across all datasets. Code and data are available here: https://github.com/maximek3/e-ViL.
****

Universal Cross-Domain Retrieval: Generalizing Across Classes and Domains
Soumava Paul, Titir Dutta, Soma Biswas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12056-12064
In this work, for the first time, we address the problem of universal cross-domain retrieval, where the test data can belong to classes or domains which are unseen during training. Due to dynamically increasing number of categories and practical constraint of training on every possible domain, which requires large amounts of data, generalizing to both unseen classes and domains is important. Towards that goal, we propose SnMpNet (Semantic Neighbourhood and Mixture Prediction Network), which incorporates two novel losses to account for the unseen classes and domains encountered during testing. Specifically, we introduce a novel Semantic Neighborhood loss to bridge the knowledge gap between seen and unseen classes and ensure that the latent space embedding of the unseen classes is semantically meaningful with respect to its neighboring classes. We also introduce a mix-up based supervision at image-level as well as semantic-level of the data for training with the Mixture Prediction loss, which helps in efficient retrieval when the query belongs to an unseen domain. These losses are incorporated on the SE-ResNet50 backbone to obtain SnMpNet. Extensive experiments on two large-scale datasets, Sketchy Extended and DomainNet, and thorough comparisons with state-of-the-art justify the effectiveness of the proposed model.
****

Learning Unsupervised Metaformer for Anomaly Detection
Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, Tyng-Luh Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4369-4378
Anomaly detection (AD) aims to address the task of classification or localization of image anomalies. This paper addresses two pivotal issues of reconstruction-

based approaches to AD in images, namely, model adaptation and reconstruction gap. The former generalizes an AD model to tackling a broad range of object categories, while the latter provides useful clues for localizing abnormal regions. At the core of our method is an unsupervised universal model, termed as Metaformer, which leverages both meta-learned model parameters to achieve high model adaptation capability and instance-aware attention to emphasize the focal regions for localizing abnormal regions, i.e., to explore the reconstruction gap at those regions of interest. We justify the effectiveness of our method with SOTA results on the MVTec AD dataset of industrial images and highlight the adaptation flexibility of the universal Metaformer with multi-class and few-shot scenarios.
****

Robust Object Detection via Instance-Level Temporal Cycle Confusion
Xin Wang, Thomas E. Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E. Gonzalez, Trevor Darrell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9143-9152
Building reliable object detectors that are robust to domain shifts, such as various changes in context, viewpoint, and object appearances, is critical for real-world applications. In this work, we study the effectiveness of auxiliary self-supervised tasks to improve the out-of-distribution generalization of object detectors. Inspired by the principle of maximum entropy, we introduce a novel self-supervised task, instance-level temporal cycle confusion (CycConf), which operates on the region features of the object detectors. For each object, the task is to find the most different object proposals in the adjacent frame in a video and then cycle back to itself for self-supervision. CycConf encourages the object detector to explore invariant structures across instances under various motions, which leads to improved model robustness in unseen domains at test time. We observe consistent out-of-domain performance improvements when training object detectors in tandem with self-supervised tasks on various domain adaptation benchmarks with static images (Cityscapes, Foggy Cityscapes, Sim10K) and large-scale video datasets (BDD100K and Waymo open data). The code and models are released at https://xinw.ai/cyc-conf.
****

HighlightMe: Detecting Highlights From Human-Centric Videos
Uttaran Bhattacharya, Gang Wu, Stefano Petrangeli, Viswanathan Swaminathan, Dinesh Manocha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8157-8167
We present a domain- and user-preference-agnostic approach to detect highlightable excerpts from human-centric videos. Our method works on the graph-based representation of multiple observable human-centric modalities in the videos, such as poses and faces. We use an autoencoder network equipped with spatial-temporal graph convolutions to detect human activities and interactions based on these modalities. We train our network to map the activity- and interaction-based latent structural representations of the different modalities to per-frame highlight scores based on the representativeness of the frames. We use these scores to compute which frames to highlight and stitch contiguous frames to produce the excerpts. We train our network on the large-scale AVA-Kinetics action dataset and evaluate it on four benchmark video highlight datasets: DSH, TVSum, PHD^2, and SumMe. We observe a 4-12% improvement in the mean average precision of matching the human-annotated highlights over state-of-the-art methods in these datasets, without requiring any user-provided preferences or dataset-specific fine-tuning.
****

Procedure Planning in Instructional Videos via Contextual Modeling and Model-Based Policy Learning
Jing Bi, Jiebo Luo, Chenliang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15611-15620
Learning new skills by observing humans' behaviors is an essential capability of AI. In this work, we leverage instructional videos to study humans' decision-making processes, focusing on learning a model to plan goal-directed actions in real-life videos. In contrast to conventional action recognition, goal-directed actions are based on expectations of their outcomes requiring causal knowledge of

potential consequences of actions. Thus, integrating the environment structure with goals is critical for solving this task. Previous works learn a single world model will fail to distinguish various tasks, resulting in an ambiguous latent space; planning through it will gradually neglect the desired outcomes since the global information of the future goal degrades quickly as the procedure evolves. We address these limitations with a new formulation of procedure planning and propose novel algorithms to model human behaviors through Bayesian Inference and model-based Imitation Learning. Experiments conducted on real-world instructional videos show that our method can achieve state-of-the-art performance in reaching the indicated goals. Furthermore, the learned contextual information presents interesting features for planning in a latent space.
****

Variable-Rate Deep Image Compression Through Spatially-Adaptive Feature Transform

Myungseo Song, Jinyoung Choi, Bohyung Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2380-2389

We propose a versatile deep image compression network based on Spatial Feature Transform (SFT), which takes a source image and a corresponding quality map as inputs and produce a compressed image with variable rates. Our model covers a wide range of compression rates using a single model, which is controlled by arbitrary pixel-wise quality maps. In addition, the proposed framework allows us to perform task-aware image compressions for various tasks, e.g., classification, by efficiently estimating optimized quality maps specific to target tasks for our encoding network. This is even possible with a pretrained network without learning separate models for individual tasks. Our algorithm achieves outstanding rate-distortion trade-off compared to the approaches based on multiple models that are optimized separately for several different target rates. At the same level of compression, the proposed approach successfully improves performance on image classification and text region quality preservation via task-aware quality map estimation without additional model training. The code is available at the project website https://github.com/micmic123/QmapCompression.
****

DeePSD: Automatic Deep Skinning and Pose Space Deformation for 3D Garment Animation

Hugo Bertiche, Meysam Madadi, Emilio Tylson, Sergio Escalera; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5471-5480

We present a novel solution to the garment animation problem through deep learning. Our contribution allows animating any template outfit with arbitrary topology and geometric complexity. Recent works develop models for garment edition, resizing and animation at the same time by leveraging the support body model (encoding garments as body homotopies). This leads to complex engineering solutions that suffer from scalability, applicability and compatibility. By limiting our scope to garment animation only, we are able to propose a simple model that can animate any outfit, independently of its topology, vertex order or connectivity. Our proposed architecture maps outfits to animated 3D models into the standard format for 3D animation (blend weights and blend shapes matrices), automatically providing of compatibility with any graphics engine. We also propose a methodology to complement supervised learning with an unsupervised physically based learning that implicitly solves collisions and enhances cloth quality.
****

Structured Outdoor Architecture Reconstruction by Exploration and Classification

Fuyang Zhang, Xiang Xu, Nelson Nauata, Yasutaka Furukawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12427-12435

This paper presents an explore-and-classify framework for structured architectural reconstruction from aerial image. Starting from a potentially imperfect building reconstruction by an existing algorithm, our approach 1) explores the space of building models by modifying the reconstruction via heuristic actions; 2) learns to classify the correctness of building models while generating classification labels based on the ground-truth; and 3) repeat. At test time, we iterate exp

loration and classification, seeking for a result with the best classification s
core. We evaluate the approach using initial reconstructions by two baselines an
d two state-of-the-art reconstruction algorithms. Qualitative and quantitative e
valuations demonstrate that our approach consistently improves the reconstructio
n quality from every initial reconstruction.
****

MG-GAN: A Multi-Generator Model Preventing Out-of-Distribution Samples in Pedest
rian Trajectory Prediction
Patrick Dendorfer, Sven Elflein, Laura Leal-Taixé; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 13158-13167
Pedestrian trajectory prediction is challenging due to its uncertain and multimo
dal nature. While generative adversarial networks can learn a distribution over
future trajectories, they tend to predict out-of-distribution samples when the d
istribution of future trajectories is a mixture of multiple, possibly disconnect
ed modes. To address this issue, we propose a multi-generator model for pedestri
an trajectory prediction. Each generator specializes in learning a distribution
over trajectories routing towards one of the primary modes in the scene, while a
 second network learns a categorical distribution over these generators, conditi
oned on the dynamics and scene input. This architecture allows us to effectively
 sample from specialized generators and to significantly reduce the out-of-distr
ibution samples compared to single generator methods.
****

Rethinking Coarse-To-Fine Approach in Single Image Deblurring
Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, Sung-Jea Ko; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4
641-4650
Coarse-to-fine strategies have been extensively used for the architecture design
 of single image deblurring networks. Conventional methods typically stack sub-n
etworks with multi-scale input images and gradually improve sharpness of images
from the bottom sub-network to the top sub-network, yielding inevitably high com
putational costs. Toward a fast and accurate deblurring network design, we revis
it the coarse-to-fine strategy and present a multi-input multi-output U-net (MIM
O-UNet). The MIMO-UNet has three distinct features. First, the single encoder of
 the MIMO-UNet takes multi-scale input images to ease the difficulty of training
. Second, the single decoder of the MIMO-UNet outputs multiple deblurred images
with different scales to mimic multi-cascaded U-nets using a single U-shaped net
work. Last, asymmetric feature fusion is introduced to merge multi-scale feature
s in an efficient manner. Extensive experiments on the GoPro and RealBlur datase
ts demonstrate that the proposed network outperforms the state-of-the-art method
s in terms of both accuracy and computational complexity. Source code is availab
le for research purposes at https://github.com/chosj95/MIMO-UNet.
****

Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation
Rawal Khirodkar, Visesh Chari, Amit Agrawal, Ambrish Tyagi; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3122-3131
A key assumption of top-down human pose estimation approaches is their expectati
on of having a single person/instance present in the input bounding box. This of
ten leads to failures in crowded scenes with occlusions. We propose a novel solu
tion to overcome the limitations of this fundamental assumption. Our Multi-Insta
nce Pose Network (MIPNet) allows for predicting multiple 2D pose instances withi
n a given bounding box. We introduce a Multi-Instance Modulation Block (MIMB) th
at can adaptively modulate channel-wise feature responses for each instance and
is parameter efficient. We demonstrate the efficacy of our approach by evaluatin
g on COCO, CrowdPose, and OCHuman datasets. Specifically, we achieve 70.0 AP on
CrowdPose and 42.5 AP on OCHuman test sets, a significant improvement of 2.4 AP
and 6.5 AP over the prior art, respectively. When using ground truth bounding bo
xes for inference, MIPNet achieves an improvement of 0.7 AP on COCO, 0.9 AP on C
rowdPose, and 9.1 AP on OCHuman validation sets compared to HRNet. Interestingly
, when fewer, high confidence bounding boxes are used, HRNet's performance degra
des (by 5 AP) on OCHuman, whereas MIPNet maintains a relatively stable performan

ce (drop of 1 AP) for the same inputs.
****

Rational Polynomial Camera Model Warping for Deep Learning Based Satellite Multi-View Stereo Matching
Jian Gao, Jin Liu, Shunping Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6148-6157
Satellite multi-view stereo (MVS) imagery is particularly suited for large-scale Earth surface reconstruction. Differing from the perspective camera model (pin-hole model) that is commonly used for close-range and aerial cameras, the cubic rational polynomial camera (RPC) model is the mainstream model for push-broom linear-array satellite cameras. However, the homography warping used in the prevailing learning based MVS methods is only applicable to pin-hole cameras. In order to apply the SOTA learning based MVS technology to the satellite MVS taskfor large-scale Earth surface reconstruction, RPC warping should be considered. In this work, we propose, for the first time, a rigorous RPC warping module. The rational polynomial coefficients are recorded as a tensor, and the RPC warping is formulated as a series of tensor transformations. Based on the RPC warping, we propose the deep learning based satellite MVS (SatMVS) framework for large-scale and wide depth range Earth surface reconstruction. We also introduce a large-scale satellite image dataset consisting of 519 5120x5120 images, which we call the TLC SatMVS dataset. The satellite images were acquired from a three-line camera (TLC) that catches triple-view images simultaneously, forming a valuable supplement to the existing open-source WorldView-3 datasets with single-scanline images. Experiments show that the proposed RPC warping module and the SatMVS framework can achieve a superior reconstruction accuracy compared to the pin-hole fitting method and conventional MVS methods. Code and data are available at https://github.com/WHU-GPCV/SatMVS.
****

A New Journey From SDRTV to HDRTV
Xiangyu Chen, Zhengwen Zhang, Jimmy S. Ren, Lynhoo Tian, Yu Qiao, Chao Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4500-4509
Nowadays modern displays are capable to render video content with high dynamic range (HDR) and wide color gamut (WCG). However, most available resources are still in standard dynamic range (SDR). Therefore, there is an urgent demand to transform existing SDR-TV contents into their HDR-TV versions. In this paper, we conduct an analysis of SDRTV-to-HDRTV task by modeling the formation of SDRTV/HDRTV content. Base on the analysis, we propose a three-step solution pipeline including adaptive global color mapping, local enhancement and highlight generation. Moreover, the above analysis inspires us to present a lightweight network that utilizes global statistics as guidance to conduct image-adaptive color mapping. In addition, we construct a dataset using HDR videos in HDR10 standard, named HDRTV1K, and select five metrics to evaluate the results of SDRTV-to-HDRTV algorithms. Furthermore, our final results achieve state-of-the-art performance in quantitative comparisons and visual quality. The code and dataset are available at https://github.com/chxy95/HDRTVNet.
****

Point-Set Distances for Learning Representations of 3D Point Clouds
Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, Binh-Son Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10478-10487
Learning an effective representation of 3D point clouds requires a good metric to measure the discrepancy between two 3D point sets, which is non-trivial due to their irregularity. Most of the previous works resort to using the Chamfer discrepancy or Earth Mover's distance, but those metrics are either ineffective in measuring the differences between point clouds or computationally expensive. In this paper, we conduct a systematic study with extensive experiments on distance metrics for 3D point clouds. From this study, we propose to use sliced Wasserstein distance and its variants for learning representations of 3D point clouds. In addition, we introduce a new algorithm to estimate sliced Wasserstein distance

that guarantees that the estimated value is close enough to the true one. Experiments show that the sliced Wasserstein distance and its variants allow the neural network to learn a more efficient representation compared to the Chamfer discrepancy. We demonstrate the efficiency of the sliced Wasserstein metric and its variants on several tasks in 3D computer vision including training a point cloud autoencoder, generative modeling, transfer learning, and point cloud registration.

****

ELLIPSDF: Joint Object Pose and Shape Optimization With a Bi-Level Ellipsoid and Signed Distance Function Description

Mo Shan, Qiaojun Feng, You-Yi Jau, Nikolay Atanasov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5946-5955

Autonomous systems need to understand the semantics and geometry of their surroundings in order to comprehend and safely execute object-level task specifications. This paper proposes an expressive yet compact model for joint object pose and shape optimization, and an associated optimization algorithm to infer an object-level map from multi-view RGB-D camera observations. The model is expressive because it captures the identities, positions, orientations, and shapes of objects in the environment. It is compact because it relies on a low-dimensional latent representation of implicit object shape, allowing onboard storage of large multi-category object maps. Different from other works that rely on a single object representation format, our approach has a bi-level object model that captures both the coarse level scale as well as the fine level shape details. Our approach is evaluated on the large-scale real-world ScanNet dataset and compared against state-of-the-art methods.

****

ARCH++: Animation-Ready Clothed Human Reconstruction Revisited

Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, Tony Tung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11046-11056

We present ARCH++, an image-based method to reconstruct 3D avatars with arbitrary clothing styles. Our reconstructed avatars are animation-ready and highly realistic, in both the visible regions from input views and the unseen regions. While prior work shows great promise of reconstructing animatable clothed humans with various topologies, we observe that there exist fundamental limitations resulting in sub-optimal reconstruction quality. In this paper, we revisit the major steps of image-based avatar reconstruction and address the limitations with ARCH++. First, we introduce an end-to-end point based geometry encoder to better describe the semantics of the underlying 3D human body, in replacement of previous hand-crafted features. Second, in order to address the occupancy ambiguity caused by topological changes of clothed humans in the canonical pose, we propose a co-supervising framework with cross-space consistency to jointly estimate the occupancy in both the posed and canonical spaces. Last, we use image-to-image translation networks to further refine detailed geometry and texture on the reconstructed surface, which improves the fidelity and consistency across arbitrary viewpoints. In the experiments, we demonstrate improvements over the state of the art on both public benchmarks and user studies in reconstruction quality and realism.

****

Vision-Language Transformer and Query Generation for Referring Segmentation

Henghui Ding, Chang Liu, Suchen Wang, Xudong Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16321-16330

In this work, we address the challenging task of referring segmentation. The query expression in referring segmentation typically indicates the target object by describing its relationship with others. Therefore, to find the target one among all instances in the image, the model must have a holistic understanding of the whole image. To achieve this, we reformulate referring segmentation as a direct attention problem: finding the region in the image where the query language expression is most attended to. We introduce transformer and multi-head attention to build a network with an encoder-decoder attention mechanism architecture that

"queries" the given image with the language expression. Furthermore, we propose a Query Generation Module, which produces multiple sets of queries with different attention weights that represent the diversified comprehensions of the language expression from different aspects. At the same time, to find the best way from these diversified comprehensions based on visual clues, we further propose a Query Balance Module to adaptively select the output features of these queries for a better mask generation. Without bells and whistles, our approach is light-weight and achieves new state-of-the-art performance consistently on three referring segmentation datasets, RefCOCO, RefCOCO+, and G-Ref. Our code is available at https://github.com/henghuiding/Vision-Language-Transformer.
****

Semantically Coherent Out-of-Distribution Detection
Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8301-8309
Current out-of-distribution (OOD) detection benchmarks are commonly built by defining one dataset as in-distribution (ID) and all others as OOD. However, these benchmarks unfortunately introduce some unwanted and impractical goals, e.g., to perfectly distinguish CIFAR dogs from ImageNet dogs, even though they have the same semantics and negligible covariate shifts. These unrealistic goals will result in an extremely narrow range of model capabilities, greatly limiting their use in real applications. To overcome these drawbacks, we re-design the benchmarks and propose the semantically coherent out-of-distribution detection (SC-OOD). On the SC-OOD benchmarks, existing methods suffer from large performance degradation, suggesting that they are extremely sensitive to low-level discrepancy between data sources while ignoring their inherent semantics. To develop an effective SC-OOD detection approach, we leverage an external un- labeled set and design a concise framework featured by unsupervised dual grouping (UDG) for the joint modeling of ID and OOD data. The proposed UDG can not only enrich the semantic knowledge of the model by exploiting unlabeled data in an unsupervised manner but also distinguish ID/OOD samples to enhance ID classification and OOD detection tasks simultaneously. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on SC-OOD benchmarks. Code and benchmarks are provided on our project page: https://jingkang50.github.io/projects/scood.
****

SCOUTER: Slot Attention-Based Classifier for Explainable Image Recognition
Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1046-1055
Explainable artificial intelligence has been gaining attention in the past few years. However, most existing methods are based on gradients or intermediate features, which are not directly involved in the decision-making process of the classifier. In this paper, we propose a slot attention-based classifier called SCOUTER for transparent yet accurate classification. Two major differences from other attention-based methods include: (a) SCOUTER's explanation is involved in the final confidence for each category, offering more intuitive interpretation, and (b) all the categories have their corresponding positive or negative explanation, which tells "why the image is of a certain category" or "why the image is not of a certain category." We design a new loss tailored for SCOUTER that controls the model's behavior to switch between positive and negative explanations, as well as the size of explanatory regions. Experimental results show that SCOUTER can give better visual explanations in terms of various metrics while keeping good accuracy on small and medium-sized datasets.
****

RetrievalFuse: Neural 3D Scene Reconstruction With a Database
Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, Angela Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12568-12577
3D reconstruction of large scenes is a challenging problem due to the high-complexity nature of the solution space, in particular for generative neural networks

. In contrast to traditional generative learned models which encode the full gen
erative process into a neural network and can struggle with maintaining local de
tails at the scene level, we introduce a new method that directly leverages scen
e geometry from the training database. First, we learn to synthesize an initial
estimate for a 3D scene, constructed by retrieving a top-k set of volumetric chu
nks from the scene database. These candidates are then refined to a final scene
generation with an attention-based refinement that can effectively select the mo
st consistent set of geometry from the candidates and combine them together to c
reate an output scene, facilitating transfer of coherent structures and local de
tail from train scene geometry. We demonstrate our neural scene reconstruction w
ith a database for the tasks of 3D super-resolution and surface reconstruction f
rom sparse point clouds, showing that our approach enables generation of more co
herent, accurate 3D scenes, improving on average by over 8% in IoU over state-of
-the-art scene reconstruction.
****

Spatio-Temporal Self-Supervised Representation Learning for 3D Point Clouds
Siyuan Huang, Yichen Xie, Song-Chun Zhu, Yixin Zhu; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2021, pp. 6535-6545
To date, various 3D scene understanding tasks still lack practical and generaliz
able pre-trained models, primarily due to the intricate nature of 3D scene under
standing tasks and their immerse variations due to camera views, lighting, occlu
sions, etc. In this paper, we tackle this immanent challenge by introducing a sp
atio-temporal representation learning (STRL) framework, capable of learning from
 unlabeled 3D point clouds in a self-supervised fashion. Inspired by how infants
 learn from visual data in-the-wild, we explore the rich spatio-temporal cues de
rived from the 3D data. Specifically, STRL takes two temporal-correlated frames
from a 3D point cloud sequence as the input, transforms it with spatial data aug
mentation, and learns the invariant representation self-supervisedly. To corrobo
rate the efficacy of STRL, we conduct extensive experiments on synthetic, indoor
, and outdoor datasets. Experimental results demonstrate that, compared with sup
ervised learning methods, the learned self-supervised representation facilitates
 various models to attain comparable or even better performances while capable o
f generalizing pre-trained models to downstream tasks, including 3D shape classi
fication, 3D object detection, and 3D semantic segmentation. Moreover, spatio-te
mporal contextual cues embedded in 3D point clouds significantly improve the lea
rned representations.
****

Learning To Know Where To See: A Visibility-Aware Approach for Occluded Person R
e-Identification
Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying
-Cong Chen, Wei-Shi Zheng; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2021, pp. 11885-11894
Person re-identification (ReID) has gained an impressive progress in recent year
s. However, the occlusion is still a common and challenging problem for recent R
eID methods. Several mainstream methods utilize extra cues (e.g., human pose inf
ormation) to distinguish human parts from obstacles to alleviate the occlusion p
roblem. Although achieving inspiring progress, these methods severely rely on th
e fine-grained extra cues, and are sensitive to the estimation error in the extr
a cues. In this paper, we show that existing methods may degrade if the extra in
formation is sparse or noisy. Thus we propose a simple yet effective method that
 is robust to sparse and noisy pose information. This is achieved by discretizin
g pose information to the visibility label of body parts, so as to suppress the
influence of occluded regions. We show in our experiments that leveraging pose i
nformation in this way is more effective and robust. Besides, our method can be
embedded into most person ReID models easily. Extensive experiments validate the
 effectiveness of our model on common occluded person ReID datasets.
****

Focal Frequency Loss for Image Reconstruction and Synthesis
Liming Jiang, Bo Dai, Wayne Wu, Chen Change Loy; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 13919-13929

Image reconstruction and synthesis have witnessed remarkable progress thanks to the development of generative models. Nonetheless, gaps could still exist between the real and generated images, especially in the frequency domain. In this study, we show that narrowing gaps in the frequency domain can ameliorate image reconstruction and synthesis quality further. We propose a novel focal frequency loss, which allows a model to adaptively focus on frequency components that are hard to synthesize by down-weighting the easy ones. This objective function is complementary to existing spatial losses, offering great impedance against the loss of important frequency information due to the inherent bias of neural networks. We demonstrate the versatility and effectiveness of focal frequency loss to improve popular models, such as VAE, pix2pix, and SPADE, in both perceptual quality and quantitative performance. We further show its potential on StyleGAN2.
****

Calibrating Concepts and Operations: Towards Symbolic Reasoning on Real Images
Zhuowan Li, Elias Stengel-Eskin, Yixiao Zhang, Cihang Xie, Quan Hung Tran, Benjamin Van Durme, Alan Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14910-14919
While neural symbolic methods demonstrate impressive performance in visual question answering on synthetic images, their performance suffers on real images. We identify that the long-tail distribution of visual concepts and unequal importance of reasoning steps in real data are the two key obstacles that limit the models' real-world potentials. To address these challenges, we propose a new paradigm, Calibrating Concepts and Operations (CCO), which enables neural symbolic models to capture underlying data characteristics and to reason with hierarchical importance. Specifically, we introduce an executor with learnable concept embedding magnitudes for handling distribution imbalance, and an operation calibrator for highlighting important operations and suppressing redundant ones. Our experiments show CCO substantially boosts the performance of neural symbolic methods on real images. By evaluating models on the real world dataset GQA, CCO helps the neural symbolic method NSCL outperforms its vanilla counterpart by 9.1% (from 47.0% to 56.1%); this result also largely reduces the performance gap between symbolic and non-symbolic methods. Additionally, we create a perturbed test set for better understanding and analyzing model performance on real images. Code is available at https://lizw14.github.io/project/ccosr.
****

Vision-Language Navigation With Random Environmental Mixup
Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, Yi-Dong Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1644-1654
Vision-language Navigation (VLN) task requires an agent to perceive both the visual scene and natural language and navigate step-by-step. Large data bias makes the VLN task challenging, which is caused by the disparity ratio between small data scale and large navigation space. Previous works have proposed many data augmentation methods to reduce data bias. However, these works do not explicitly reduce the data bias across different house scenes. Therefore, the agent would be overfitting to the seen scenes and perform navigation poorly in the unseen scenes. To tackle this problem, we propose the random environmental mixup (REM) method, which generates augmentation data in cross-connected house scenes. This method consists of three steps: 1) we select the key viewpoints according to the room connection graph for each scene in the training split; 2) we cross-connect the key views of different scenes to construct augmented scenes; 3) we generate augmentation data triplets (environment, path, instruction) in the cross-connected scenes. Our experiments prove that the augmentation data helps the agent reduce its performance gap between the seen and unseen environment and improve its performance, making our model be the best existing approach on the standard benchmark.
****

Object Tracking by Jointly Exploiting Frame and Event Domain
Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, Bo Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.

13043-13052
Inspired by the complementarity between conventional frame-based and bio-inspired event-based cameras, we propose a multi-modal based approach to fuse visual cues from the frame- and event-domain to enhance the single object tracking performance, especially in degraded conditions (e.g., scenes with high dynamic range, low light, and fast-motion objects). The proposed approach can effectively and adaptively combine meaningful information from both domains. Our approach's effectiveness is enforced by a novel designed cross-domain attention schemes, which can effectively enhance features based on self- and cross-domain attention schemes; The adaptiveness is guarded by a specially designed weighting scheme, which can adaptively balance the contribution of the two domains. To exploit event-based visual cues in single-object tracking, we construct a large-scale frame-event-based dataset, which we subsequently employ to train a novel frame-event fusion based model. Extensive experiments show that the proposed approach outperforms state-of-the-art frame-based tracking methods by at least 10.4% and 11.9% in terms of representative success rate and precision rate, respectively. Besides, the effectiveness of each key component of our approach is evidenced by our thorough ablation study.
****
Learning To Generate Scene Graph From Natural Language Supervision
Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, Yin Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1823-1834
Learning from image-text data has demonstrated recent success for many recognition tasks, yet is currently limited to visual features or individual visual concepts such as objects. In this paper, we propose one of the first methods that learn from image-sentence pairs to extract a graphical representation of localized objects and their relationships within an image, known as scene graph. To bridge the gap between images and texts, we leverage an off-the-shelf object detector to identify and localize object instances, match labels of detected regions to concepts parsed from captions, and thus create "pseudo" labels for learning scene graph. Further, we design a Transformer-based model to predict these "pseudo" labels via a masked token prediction task. Learning from only image-sentence pairs, our model achieves 30% relative gain over a latest method trained with human-annotated unlocalized scene graphs. Our model also shows strong results for weakly and fully supervised scene graph generation. In addition, we explore an open-vocabulary setting for detecting scene graphs, and present the first result for open-set scene graph generation.
****
Editing Conditional Radiance Fields
Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, Bryan Russell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5773-5783
A neural radiance field (NeRF) is a scene model supporting high-quality view synthesis, optimized per scene. In this paper, we explore enabling user editing of a category-level NeRF trained on a shape category. Specifically, we propose a method for propagating coarse 2D user scribbles to the 3D space, to modify the color or shape of a local region. First, we propose a conditional radiance field that incorporates new modular network components, including a branch that is shared across object instances in the category. Observing multiple instances of the same category, our model learns underlying part semantics without any supervision, thereby allowing the propagation of coarse 2D user scribbles to the entire 3D region (e.g., chair seat) in a consistent fashion. Next, we investigate for the editing tasks which components of our network require updating. We propose a hybrid network update strategy that targets the later network components, which balances efficiency and accuracy. During user interaction, we formulate an optimization problem that both satisfies the user's constraints and preserves the original object structure. We demonstrate our approach on a variety of editing tasks over three shape datasets and show that it outperforms prior neural editing approaches. Finally, we edit the appearance and shape of a real photograph and show that the edit propagates to extrapolated novel views.

****
Global Pooling, More Than Meets the Eye: Position Information Is Encoded Channel-Wise in CNNs
Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G. Derpanis, Neil D. B. Bruce; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 793-801
In this paper, we challenge the common assumption that collapsing the spatial dimensions of a 3D (spatial-channel) tensor in a convolutional neural network (CNN) into a vector via global pooling removes all spatial information. Specifically, we demonstrate that positional information is encoded based on the ordering of the channel dimensions, while semantic information is largely not. Following this demonstration, we show the real world impact of these findings by applying them to two applications. First, we propose a simple yet effective data augmentation strategy and loss function which improves the translation invariance of a CNN's output. Second, we propose a method to efficiently determine which channels in the latent representation are responsible for (i) encoding overall position information or (ii) region-specific positions. We first show that semantic segmentation has a significant reliance on the overall position channels to make predictions. We then show for the first time that it is possible to perform a `region-specific' attack, and degrade a network's performance in a particular part of the input. We believe our findings and demonstrated applications will benefit research areas concerned with understanding the characteristics of CNNs.
****
Testing Using Privileged Information by Adapting Features With Statistical Dependence
Kwang In Kim, James Tompkin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9405-9413
Given an imperfect predictor, we exploit additional features at test time to improve the predictions made, without retraining and without knowledge of the prediction function. This scenario arises if training labels or data are proprietary, restricted, or no longer available, or if training itself is prohibitively expensive. We assume that the additional features are useful if they exhibit strong statistical dependence to the underlying perfect predictor. Then, we empirically estimate and strengthen the statistical dependence between the initial noisy predictor and the additional features via manifold denoising. As an example, we show that this approach leads to improvement in real-world visual attribute ranking.
****
CvT: Introducing Convolutions to Vision Transformers
Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 22-31
We present in this paper a new architecture, named Convolutional vision Transformer (CvT), that improves Vision Transformer (ViT) in performance and efficiency by introducing convolutions into ViT to yield the best of both designs. This is accomplished through two primary modifications: a hierarchy of Transformers containing a new convolutional token embedding, and a convolutional Trasnsformer block leveraging a convolutional projection. These changes introduce desirable properties of convolutional neural networks (CNNs) to the ViT architecture (i.e. shift, scale, and distortion invariance) while maintaining the merits of Transformers (i.e. dynamic attention, global context, and better generalization). We validate CvT by conducting extensive experiments, showing that this approach achieves state-of-the-art performance over other Vision Transformers and ResNets on ImageNet-1k, with less parameters and lower FLOPs. In addition, performance gains are maintained when pretrained on larger datasets (e.g. ImageNet-22k) and fine-tuned to downstream tasks. Finally, our results show that the positional encoding, a crucial component in existing Vision Transformers, can be safely removed in our model, simplifying the design for higher resolution vision tasks. Code will be released at https://github.com/microsoft/CvT.
****

Context-Sensitive Temporal Feature Learning for Gait Recognition
Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu , Bin Feng; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 12909-12918
Although gait recognition has drawn increasing research attention recently, it r emains challenging to learn discriminative temporal representation since the sil houette differences are quite subtle in spatial domain. Inspired by the observat ion that humans can distinguish gaits of different subjects by adaptively focusi ng on temporal sequences with different time scales, we propose a context-sensit ive temporal feature learning (CSTL) network in this paper, which aggregates tem poral features in three scales to obtain motion representation according to the temporal contextual information. Specifically, CSTL introduces relation modeling among multi-scale features to evaluate feature importances, based on which netw ork adaptively enhances more important scale and suppresses less important scale . Besides that, we propose a salient spatial feature learning (SSFL) module to t ackle the misalignment problem caused by temporal operation, e.g., temporal conv olution. SSFL recombines a frame of salient spatial features by extracting the m ost discriminative parts across the whole sequence. In this way, we achieve adap tive temporal learning and salient spatial mining simultaneously. Extensive expe riments conducted on two datasets demonstrate the state-of-the-art performance. On CASIA-B dataset, we achieve rank-1 accuracies of 98.0%, 95.4% and 87.0% under normal walking, bag-carrying and coat-wearing conditions. On OU-MVLP dataset, w e achieve rank-1 accuracy of 90.2%. The source code will be published at https:/ /github.com/OliverHxh/CSTL.
****

Pseudo-Mask Matters in Weakly-Supervised Semantic Segmentation
Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, Wayne Zhang; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6964-6973
Most weakly supervised semantic segmentation (WSSS) methods follow the pipeline that generates pseudo-masks initially and trains the segmentation model with the pseudo-masks in fully supervised manner after. However, we find some matters re lated to the pseudo-masks, including high quality pseudo-masks generation from c lass activation maps (CAMs), and training with noisy pseudo-mask supervision. Fo r these matters, we propose the following designs to push the performance to new state-of-art: (i) Coefficient of Variation Smoothing to smooth the CAMs adaptiv ely; (ii) Proportional Pseudo-mask Generation to project the expanded CAMs to ps eudo-mask based on a new metric indicating the importance of each class on each location, instead of the scores trained from binary classifiers. (iii) Pretended Under-Fitting strategy to suppress the influence of noise in pseudo-mask; (iv) Cyclic Pseudo-mask to boost the pseudo-masks during training of fully supervised semantic segmentation (FSSS). Experiments based on our methods achieve new stat e-of-art results on two changeling weakly supervised semantic segmentation datas ets, pushing the mIoU to 70.0% and 40.2% on PAS-CAL VOC 2012 and MS COCO 2014 re spectively. Codes including segmentation framework are released at https://githu b.com/Eli-YiLi/PMM
****

COTR: Correspondence Transformer for Matching Across Images
Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, Kwang Moo Yi; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p p. 6207-6217
We propose a novel framework for finding correspondences in images based on a de ep neural network that, given two images and a query point in one of them, finds its correspondence in the other. By doing so, one has the option to query only the points of interest and retrieve sparse correspondences, or to query all poin ts in an image and obtain dense mappings. Importantly, in order to capture both local and global priors, and to let our model relate between image regions using the most relevant among said priors, we realize our network using a transformer . At inference time, we apply our correspondence network by recursively zooming in around the estimates, yielding a multi-scale pipeline able to provide highly-accurate correspondences. Our method significantly outperforms the state-of-the-

art on both sparse and dense correspondence problems on multiple datasets and ta
sks, ranging from wide-baseline stereo to optical flow, without any retraining f
or a specific dataset.
****

CoMatch: Semi-Supervised Learning With Contrastive Graph Regularization
Junnan Li, Caiming Xiong, Steven C.H. Hoi; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 9475-9484
Semi-supervised learning has been an effective paradigm for leveraging unlabeled
 data to reduce the reliance on labeled data. We propose CoMatch, a new semi-sup
ervised learning method that unifies dominant approaches and addresses their lim
itations. CoMatch jointly learns two representations of the training data, their
 class probabilities and low-dimensional embeddings. The two representations int
eract with each other to jointly evolve. The embeddings impose a smoothness cons
traint on the class probabilities to improve the pseudo-labels, whereas the pseu
do-labels regularize the structure of the embeddings through graph-based contras
tive learning. CoMatch achieves state-of-the-art performance on multiple dataset
s. It achieves substantial accuracy improvements on the label-scarce CIFAR-10 an
d STL-10. On ImageNet with 1% labels, CoMatch achieves a top-1 accuracy of 66.0%
, outperforming FixMatch by 12.6%. Furthermore, CoMatch achieves better represen
tation learning performance on downstream tasks, outperforming both supervised l
earning and self-supervised learning. Code and pre-trained models are available
at https://github.com/salesforce/CoMatch/.
****

End-to-End Semi-Supervised Object Detection With Soft Teacher
Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang B
ai, Zicheng Liu; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2021, pp. 3060-3069
Previous pseudo-label approaches for semi-supervised object detection typically
 follow a multi-stage schema, with the first stage to train an initial detector o
n a few labeled data, followed by the pseudo labeling and re-training stage on u
nlabeled data. These multi-stage methods complicate the training, and also hinde
r the use of improved detectors for more accurate pseudo-labeling. In this paper
, we propose an end-to-end approach to simultaneously improve the detector and p
seudo labels gradually for semi-supervised object detection. The pseudo labels a
re generated on the fly by a teacher model which is an aggregated version of the
 student detector at different steps. As the detector becomes stronger during th
e training, the teacher detector's performance improves and the pseudo labels te
nd to be more accurate, which further benefits the detector training. Within the
 end-to-end training, we present two simple yet effective techniques: weigh the
classification loss of unlabeled images through soft teacher and select reliable
 pseudo boxes for regression through box jittering. Experimentally, the proposed
 approach outperforms the state-of-the-art methods by a large margin on MS-COCO
benchmark by using Faster R-CNN with ResNet-50 and FPN, reaching 20.5 mAP, 30.7
mAP and 34.0 mAP with 1%, 5%, 10% labeled data, respectively. Moreover, the prop
osed approach also proves to improve this detector trained on the COCO full set
by +1.8 mAP by leveraging additional unlabelled data of COCO, achieving 42.7 mAP
.
****

Zen-NAS: A Zero-Shot NAS for High-Performance Image Recognition
Ming Lin, Pichao Wang, Zhenhong Sun, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, Ron
g Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2021, pp. 347-356
Accuracy predictor is a key component in Neural Architecture Search (NAS) for ra
nking architectures. Building a high-quality accuracy predictor usually costs en
ormous computation. To address this issue, instead of using an accuracy predicto
r, we propose a novel zero-shot index dubbed Zen-Score to rank the architectures
. The Zen-Score represents the network expressivity and positively correlates wi
th the model accuracy. The calculation of Zen-Score only takes a few forward inf
erences through a randomly initialized network, without training network paramet
ers. Built upon the Zen-Score, we further propose a new NAS algorithm, termed as

Zen-NAS, by maximizing the Zen-Score of the target network under given inferenc e budgets. Within less than half GPU day, Zen-NAS is able to directly search hig h performance architectures in a data-free style. Comparing with previous NAS me thods, the proposed Zen-NAS is magnitude times faster on multiple server-side an d mobile-side GPU platforms with state-of-the-art accuracy on ImageNet. Searchin g and training code as well as pre-trained models are available from https://git hub.com/idstcv/ZenNAS.

****

Virtual Light Transport Matrices for Non-Line-of-Sight Imaging

Julio Marco, Adrian Jarabo, Ji Hyun Nam, Xiaochun Liu, Miguel Ángel Cosculluela, Andreas Velten, Diego Gutierrez; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2021, pp. 2440-2449

The light transport matrix (LTM) is an instrumental tool in line-of-sight (LOS) imaging, describing how light interacts with the scene and enabling applications such as relighting or separation of illumination components. We introduce a fra mework to estimate the LTM of non-line-of-sight (NLOS) scenarios, coupling recen t virtual forward light propagation models for NLOS imaging with the LOS light t ransport equation. We design computational projector-camera setups, and use thes e virtual imaging systems to estimate the transport matrix of hidden scenes. We introduce the specific illumination functions to compute the different elements of the matrix, overcoming the challenging wide-aperture conditions of NLOS setup s. Our NLOS light transport matrix allows us to (re)illuminate specific location s of a hidden scene, and separate direct, first-order indirect, and higher-order indirect illumination of complex cluttered hidden scenes, similar to existing L OS techniques.

****

DecentLaM: Decentralized Momentum SGD for Large-Batch Deep Training

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, Wotao Y in; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 3029-3039

The scale of deep learning nowadays calls for efficient distributed training alg orithms. Decentralized momentum SGD (DmSGD), in which each node averages only wi th its neighbors, is more communication efficient than vanilla Parallel momentum SGD that incurs global average across all computing nodes. On the other hand, t he large-batch training has been demonstrated critical to achieve runtime speedu p. This motivates us to investigate how DmSGD performs in the large-batch scenar io. In this work, we find the momentum term can amplify the inconsistency bias in DmSGD. Such bias becomes more evident as batch-size grows large and hence res ults in severe performance degradation. We next propose DecentLaM, a novel decen tralized large-batch momentum SGD to remove the momentum-incurred bias. The conv ergence rate for both strongly convex and non-convex scenarios is established. O ur theoretical results justify the superiority of DecentLaM to DmSGD especially in the large-batch scenario. Experimental results on a a variety of computer vis ion tasks and models show that DecentLaM promises both efficient and high-qualit y training.

****

Video Object Segmentation With Dynamic Memory Networks and Adaptive Object Align ment

Shuxian Liang, Xu Shen, Jianqiang Huang, Xian-Sheng Hua; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2021, pp. 8065-8074

In this paper, we propose a novel solution for object-matching based semi-superv ised video object segmentation, where the target object masks in the first frame are provided. Existing object-matching based methods focus on the matching betw een the raw object features of the current frame and the first/previous frames. However, two issues are still not solved by these object-matching based methods. As the appearance of the video object changes drastically over time, 1) unseen parts/details of the object present in the current frame, resulting in incomplet e annotation in the first annotated frame (e.g., view/scale changes). 2) even fo r the seen parts/details of the object in the current frame, their positions cha nge relatively (e.g., pose changes/camera motion), leading to a misalignment for

the object matching. To obtain the complete information of the target object, we propose a novel object-based dynamic memory network that exploits visual contents of all the past frames. To solve the misalignment problem caused by position changes of visual contents, we propose an adaptive object alignment module by incorporating a region translation function that aligns object proposals towards templates in the feature space. Our method achieves state-of-the-art results on latest benchmark datasets DAVIS 2017 (J of 81.4% and F of 87.5% on the validation set) and YouTube-VOS (the overall score of 82.7% on the validation set) with a very efficient inference time (0.16 second/frame on DAVIS 2017 validation set). Code is available at: https://github.com/liang4sx/DMN-AOA.
****

Augmented Lagrangian Adversarial Attacks
Jérôme Rony, Eric Granger, Marco Pedersoli, Ismail Ben Ayed; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7738-7747
Adversarial attack algorithms are dominated by penalty methods, which are slow in practice, or more efficient distance-customized methods, which are heavily tailored to the properties of the considered distance. We propose a white-box attack algorithm to generate minimally perturbed adversarial examples based on Augmented Lagrangian principles. We bring several algorithmic modifications, which have a crucial effect on performance. Our attack enjoys the generality of penalty methods and the computational efficiency of distance-customized algorithms, and can be readily used for a wide set of distances. We compare our attack to state-of-the-art methods on three datasets and several models, and consistently obtain competitive performances with similar or lower computational complexity.
****

Contrastive Multimodal Fusion With TupleInfoNCE
Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, Li Yi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 754-763
This paper proposes a method for representation learning of multimodal data using contrastive losses. A traditional approach is to contrast different modalities to learn the information shared between them. However, that approach could fail to learn the complementary synergies between modalities that might be useful for downstream tasks. Another approach is to concatenate all the modalities into a tuple and then contrast positive and negative tuple correspondences. However, that approach could consider only the stronger modalities while ignoring the weaker ones. To address these issues, we propose a novel contrastive learning objective, TupleInfoNCE. It contrasts tuples based not only on positive and negative correspondences, but also by composing new negative tuples using modalities describing different scenes. Training with these additional negatives encourages the learning model to examine the correspondences among modalities in the same tuple, ensuring that weak modalities are not ignored. We provide a theoretical justification based on mutual-information for why this approach works, and we propose a sample optimization algorithm to generate positive and negative samples to maximize training efficacy. We find that TupleInfoNCE significantly outperforms previous state of the arts on three different downstream tasks.
****

Deep Reparametrization of Multi-Frame Super-Resolution and Denoising
Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2460-2470
We propose a deep reparametrization of the maximum a posteriori formulation commonly employed in multi-frame image restoration tasks. Our approach is derived by introducing a learned error metric and a latent representation of the target image, which transforms the MAP objective to a deep feature space. The deep reparametrization allows us to directly model the image formation process in the latent space, and to integrate learned image priors into the prediction. Our approach thereby leverages the advantages of deep learning, while also benefiting from the principled multi-frame fusion provided by the classical MAP formulation. We validate our approach through comprehensive experiments on burst denoising and bu

rst super-resolution datasets. Our approach sets a new state-of-the-art for both tasks, demonstrating the generality and effectiveness of the proposed formulation.

****

Always Be Dreaming: A New Approach for Data-Free Class-Incremental Learning
James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, Zsolt Kira; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9374-9384

Modern computer vision applications suffer from catastrophic forgetting when incrementally learning new concepts over time. The most successful approaches to alleviate this forgetting require extensive replay of previously seen data, which is problematic when memory constraints or data legality concerns exist. In this work, we consider the high-impact problem of Data-Free Class-Incremental Learning (DFCIL), where an incremental learning agent must learn new concepts over time without storing generators or training data from past tasks. One approach for DFCIL is to replay synthetic images produced by inverting a frozen copy of the learner's classification model, but we show this approach fails for common class-incremental benchmarks when using standard distillation strategies. We diagnose the cause of this failure and propose a novel incremental distillation strategy for DFCIL, contributing a modified cross-entropy training and importance-weighted feature distillation, and show that our method results in up to a 25.1% increase in final task accuracy (absolute difference) compared to SOTA DFCIL methods for common class-incremental benchmarks. Our method even outperforms several standard replay based methods which store a coreset of images.

****

ViViT: A Video Vision Transformer
Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lu■i■, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6836-6846

We present pure-transformer based models for video classification, drawing upon the recent success of such models in image classification. Our model extracts spatio-temporal tokens from the input video, which are then encoded by a series of transformer layers. In order to handle the long sequences of tokens encountered in video, we propose several, efficient variants of our model which factorise the spatial- and temporal-dimensions of the input. Although transformer-based models are known to only be effective when large training datasets are available, we show how we can effectively regularise the model during training and leverage pretrained image models to be able to train on comparatively small datasets. We conduct thorough ablation studies, and achieve state-of-the-art results on multiple video classification benchmarks including Kinetics 400 and 600, Epic Kitchens, Something-Something v2 and Moments in Time, outperforming prior methods based on deep 3D convolutional networks. To facilitate further research, we will release code and models.

****

Generative Compositional Augmentations for Scene Graph Prediction
Boris Knyazev, Harm de Vries, C■t■lina Cangea, Graham W. Taylor, Aaron Courville, Eugene Belilovsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15827-15837

Inferring objects and their relationships from an image in the form of a scene graph is useful in many applications at the intersection of vision and language. We consider a challenging problem of compositional generalization that emerges in this task due to a long tail data distribution. Current scene graph generation models are trained on a tiny fraction of the distribution corresponding to the most frequent compositions, e.g. <cup, on, table>. However, test images might contain zero- and few-shot compositions of objects and relationships, e.g. <cup, on, surfboard>. Despite each of the object categories and the predicate (e.g. 'on') being frequent in the training data, the models often fail to properly understand such unseen or rare compositions. To improve generalization, it is natural to attempt increasing the diversity of the training distribution. However, in the graph domain this is non-trivial. To that end, we propose a method to synthesi

ze rare yet plausible scene graphs by perturbing real ones. We then propose and empirically study a model based on conditional generative adversarial networks (GANs) that allows us to generate visual features of perturbed scene graphs and l earn from them in a joint fashion. When evaluated on the Visual Genome dataset, our approach yields marginal, but consistent improvements in zero- and few-shot metrics. We analyze the limitations of our approach indicating promising directi ons for future research.
****

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery
Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, Dani Lischinski; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2085-2094
Inspired by the ability of StyleGAN to generate highly re-alistic images in a va riety of domains, much recent work hasfocused on understanding how to use the la tent spaces ofStyleGAN to manipulate generated and real images. How-ever, discov ering semantically meaningful latent manipula-tions typically involves painstaki ng human examination ofthe many degrees of freedom, or an annotated collectionof images for each desired manipulation. In this work, weexplore leveraging the po wer of recently introduced Con-trastive Language-Image Pre-training (CLIP) model s in or-der to develop a text-based interface for StyleGAN imagemanipulation tha t does not require such manual effort. Wefirst introduce an optimization scheme that utilizes a CLIP-based loss to modify an input latent vector in response to auser-provided text prompt. Next, we describe a latent map-per that infers a tex t-guided latent manipulation step fora given input image, allowing faster and mo re stable text-based manipulation. Finally, we present a method for map-ping a t ext prompts to input-agnostic directions in Style-GAN's style space, enabling in teractive text-driven imagemanipulation. Extensive results and comparisons demon -strate the effectiveness of our approaches.
****

Meta-Attack: Class-Agnostic and Model-Agnostic Physical Adversarial Attack
Weiwei Feng, Baoyuan Wu, Tianzhu Zhang, Yong Zhang, Yongdong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 77 87-7796
Modern deep neural networks are often vulnerable to adversarial examples. Most e xist attack methods focus on crafting adversarial examples in the digital domain , while only limited works study physical adversarial attack. However, it is mor e challenging to generate effective adversarial examples in the physical world d ue to many uncontrollable physical dynamics. Most current physical attack method s aim to generate robust physical adversarial examples by simulating all possibl e physical dynamics. When attacking new images or new DNN models, they require e xpensive manually efforts for simulating physical dynamics and considerable time for iteratively optimizing for each image. To tackle these issues, we propose a class-agnostic and model-agnostic physical adversarial attack model (Meta-Attac k), which is able to not only generate robust physical adversarial examples by s imulating color and shape distortions, but also generalize to attacking novel im ages and novel DNN models by accessing a few digital and physical images. To the best of our knowledge, this is the first work to formulate the physical attack as a few-shot learning problem. Here, the training task is redefined as the comp osition of a support set, a query set, and a target DNN model. Under the few- sh ot setting, we design a novel class-agnostic and model-agnostic meta-learning al gorithm to enhance the generalization ability of our method. Extensive experimen tal results on two benchmark datasets with four challenging experimental setting s verify the superior robustness and generalization of our method by comparing t o state-of-the-art physical attack methods.
****

Benchmarking Ultra-High-Definition Image Super-Resolution
Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, Björn Stenger, Wei Liu, Hongdong Li, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Co mputer Vision (ICCV), 2021, pp. 14769-14778
Increasingly, modern mobile devices allow capturing images at Ultra-High-Definit

ion (UHD) resolution, which includes 4K and 8K images. However, current single image super-resolution (SISR) methods focus on super-resolving images to ones with resolution up to high definition (HD) and ignore higher-resolution UHD images. To explore their performance on UHD images, in this paper, we first introduce two large-scale image datasets, UHDSR4K and UHDSR8K, to benchmark existing SISR methods. With 70,000 V100 GPU hours of training, we benchmark these methods on 4K and 8K resolution images under seven different settings to provide a set of baseline models. Moreover, we propose a baseline model, called Mesh Attention Network (MANet) for SISR. The MANet applies the attention mechanism in both different depths (horizontal) and different levels of receptive field (vertical). In this way, correlations among feature maps are learned, enabling the network to focus on more important features.

****

3D Local Convolutional Neural Networks for Gait Recognition

Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, Xian-Sheng Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14920-14929

The goal of gait recognition is to learn the unique spatio-temporal pattern about the human body shape from its temporal changing characteristics. As different body parts behave differently during walking, it is intuitive to model the spatio-temporal patterns of each part separately. However, existing part-based methods equally divide the feature maps of each frame into fixed horizontal stripes to get local parts. It is obvious that these stripe partition-based methods cannot accurately locate the body parts. First, different body parts can appear at the same stripe (e.g., arms and the torso), and one part can appear at different stripes in different frames (e.g., hands). Second, different body parts possess different scales, and even the same part in different frames can appear at different locations and scales. Third, different parts also exhibit distinct movement patterns (e.g., at which frame the movement starts, the position change frequency, how long it lasts). To overcome these issues, we propose novel 3D local operations as a generic family of building blocks for 3D gait recognition backbones. The proposed 3D local operations support the extraction of local 3D volumes of body parts in a sequence with adaptive spatial and temporal scales, locations and lengths. In this way, the spatio-temporal patterns of the body parts are well learned from the 3D local neighborhood in part-specific scales, locations, frequencies and lengths. Experiments demonstrate that our 3D local convolutional neural networks achieve state-of-the-art performance on popular gait datasets. Code is available at: https://github.com/yellowtownhz/3DLocalCNN.

****

Lucas-Kanade Reloaded: End-to-End Super-Resolution From Raw Image Bursts

Bruno Lecouat, Jean Ponce, Julien Mairal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2370-2379

This presentation addresses the problem of reconstructing a high-resolution image from multiple lower-resolution snapshots captured from slightly different viewpoints in space and time. Key challenges for solving this super-resolution problem include (i) aligning the input pictures with sub-pixel accuracy, (ii) handling raw (noisy) images for maximal faithfulness to native camera data, and (iii) designing/learning an image prior (regularizer) well suited to the task. We address these three challenges with a hybrid algorithm building on the insight from Wronski et al. that aliasing is an ally in this setting, with parameters that can be learned end to end, while retaining the interpretability of classical approaches to inverse problems. The effectiveness of our approach is demonstrated on synthetic and real image bursts, setting a new state of the art on several benchmarks and delivering excellent qualitative results on real raw bursts captured by smartphones and prosumer cameras.

****

Learning Better Visual Data Similarities via New Grouplet Non-Euclidean Embedding

Yanfu Zhang, Lei Luo, Wenhan Xian, Heng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9918-9927

In many computer vision problems, it is desired to learn the effective visual da ta similarity such that the prediction accuracy can be enhanced. Deep Metric Lea rning (DML) methods have been actively studied to measure the data similarity. P air-based and proxy-based losses are the two major paradigms in DML. However, pa ir-wise methods involve expensive training costs, while proxy-based methods are less accurate in characterizing the relationships between data points. In this p aper, we provide a hybrid grouplet paradigm, which inherits the accurate pair-wi se relationship in pair-based methods and the efficient training in proxy-based methods. Our method also equips a non-Euclidean space to DML, which employs a hi erarchical representation manifold. More specifically, we propose a unified grap h perspective --- different DML methods learn different local connecting pattern s between data points. Based on the graph interpretation, we construct a flexibl e subset of data points, dubbed grouplet. Our grouplet doesn't require explicit pair-wise relationships, instead, we encode the data relationships in an optimal transport problem regarding the proxies, and solve this problem via a different iable implicit layer to automatically determine the relationships. Extensive exp erimental results show that our method significantly outperforms state-of-the-ar t baselines on several benchmarks. The ablation studies also verify the effectiv eness of our method.

****

PR-GCN: A Deep Graph Convolutional Network With Point Refinement for 6D Pose Est imation

Guangyuan Zhou, Huiqun Wang, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2793-2802

RGB-D based 6D pose estimation has recently achieved remarkable progress, but st ill suffers from two major limitations: (1) ineffective representation of depth data and (2) insufficient integration of different modalities. This paper propos es a novel deep learning approach, namely Graph Convolutional Network with Point Refinement (PR-GCN), to simultaneously address the issues above in a unified wa y. It first introduces the Point Refinement Network (PRN) to polish 3D point clo uds, recovering missing parts with noise removed. Subsequently, the Multi-Modal Fusion Graph Convolutional Network (MMF-GCN) is presented to strengthen RGB-D co mbination, which captures geometry-aware inter-modality correlation through loca l information propagation in the graph convolutional network. Extensive experime nts are conducted on three widely used benchmarks, and state-of-the-art performa nce is reached. Besides, it is also shown that the proposed PRN and MMF-GCN modu les are well generalized to other frameworks.

****

Learning High-Fidelity Face Texture Completion Without Complete Face Texture

Jongyoo Kim, Jiaolong Yang, Xin Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13990-13999

For face texture completion, previous methods typically use some complete textur es captured by multiview imaging systems or 3D scanners for supervised learning. This paper deals with a new challenging problem -- learning to complete invisib le texture in a single face image without using any complete texture. We simply leverage a large corpus of face images of different subjects (e.\,g., FFHQ) to t rain a texture completion model in an unsupervised manner. To achieve this, we p ropose DSD-GAN, a novel deep neural network based method that applies two discri minators in UV map space and image space. These two discriminators work in a com plementary manner to learn both facial structures and texture details. We show t hat their combination is essential to obtain high-fidelity results. Despite the network never sees any complete facial appearance, it is able to generate compel ling full textures from single images.

****

Product Quantizer Aware Inverted Index for Scalable Nearest Neighbor Search

Haechan Noh, Taeho Kim, Jae-Pil Heo; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2021, pp. 12210-12218

The inverted index is one of the most commonly used structures for non-exhaustiv e nearest neighbor search on large-scale datasets. It allows a significant facto r of acceleration by a reduced number of distance computations with only a small

fraction of the database. In particular, the inverted index enables the product quantization (PQ) to learn their codewords in the residual vector space. The quantization error of the PQ can be substantially improved in such combination since the residual vector space is much more quantization-friendly thanks to their compact distribution compared to the original data. In this paper, we first raise an unremarked but crucial question; why the inverted index and the product quantizer are optimized separately even though they are closely related? For instance, changes on the inverted index distort the whole residual vector space. To address the raised question, we suggest a joint optimization of the coarse and fine quantizers by substituting the original objective of the coarse quantizer to end-to-end quantization distortion. Moreover, our method is generic and applicable to different combinations of coarse and fine quantizers such as inverted multi-index and optimized PQ.

****

RINDNet: Edge Detection for Discontinuity in Reflectance, Illumination, Normal and Depth

Mengyang Pu, Yaping Huang, Qingji Guan, Haibin Ling; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6879-6888

As a fundamental building block in computer vision, edges can be categorised into four types according to the discontinuity in surface-Reflectance, Illumination, surface-Normal or Depth. While great progress has been made in detecting generic or individual types of edges, it remains under-explored to comprehensively study all four edge types together. In this paper, we propose a novel neural network solution, RINDNet, to jointly detect all four types of edges. Taking into consideration the distinct attributes of each type of edges and the relationship between them, RINDNet learns effective representations for each of them and works in three stages. In stage I, RINDNet uses a common backbone to extract features shared by all edges. Then in stage II it branches to prepare discriminative features for each edge type by the corresponding decoder. In stage III, an independent decision head for each type aggregates the features from previous stages to predict the initial results. Additionally, an attention module learns attention maps for all types to capture the underlying relations between them, and these maps are combined with initial results to generate the final edge detection results. For training and evaluation, we construct the first public benchmark, BSDS-RIND, with all four types of edges carefully annotated. In our experiments, RINDNet yields promising results in comparison with state-of-the-art methods. Additional analysis is presented in supplementary material.

****

Track Without Appearance: Learn Box and Tracklet Embedding With Local and Global Motion Patterns for Vehicle Tracking

Gaoang Wang, Renshu Gu, Zuozhu Liu, Weijie Hu, Mingli Song, Jenq-Neng Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9876-9886

Vehicle tracking is an essential task in the multi-object tracking (MOT) field. A distinct characteristic in vehicle tracking is that the trajectories of vehicles are fairly smooth in both the world coordinate and the image coordinate. Hence, models that capture motion consistencies are of high necessity. However, tracking with the standalone motion-based trackers is quite challenging because targets could get lost easily due to limited information, detection error and occlusion. Leveraging appearance information to assist object re-identification could resolve this challenge to some extent. However, doing so requires extra computation while appearance information is sensitive to occlusion as well. In this paper, we try to explore the significance of motion patterns for vehicle tracking without appearance information. We propose a novel approach that tackles the association issue for long-term tracking with the exclusive fully-exploited motion information. We address the tracklet embedding issue with the proposed reconstruct-to-embed strategy based on deep graph convolutional neural networks (GCN). Comprehensive experiments on the KITTI-car tracking dataset and UA-Detrac dataset show that the proposed method, though without appearance information, could achieve competitive performance with the state-of-the-art (SOTA) trackers. The source

code will be available at https://github.com/GaoangW/LGMTracker.
****
Are We Missing Confidence in Pseudo-LiDAR Methods for Monocular 3D Object Detection?
Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Elisa Ricci; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3225-3233
Pseudo-LiDAR-based methods for monocular 3D object detection have received considerable attention in the community due to the performance gains exhibited on the KITTI3D benchmark, in particular on the commonly reported validation split. This generated a distorted impression about the superiority of Pseudo-LiDAR-based (PL-based) approaches over methods working with RGB images only. Our first contribution consists in rectifying this view by pointing out and showing experimentally that the validation results published by PL-based methods are substantially biased. The source of the bias resides in an overlap between the KITTI3D object detection validation set and the training/validation sets used to train depth predictors feeding PL-based methods. Surprisingly, the bias remains also after geographically removing the overlap. This leaves the test set as the only reliable set for comparison, where published PL-based methods do not excel. Our second contribution brings PL-based methods back up in the ranking with the design of a novel deep architecture which introduces a 3D confidence prediction module. We show that 3D confidence estimation techniques derived from RGB-only 3D detection approaches can be successfully integrated into our framework and, more importantly, that improved performance can be obtained with a newly designed 3D confidence measure, leading to state-of-the-art performance on the KITTI3D benchmark.
****
On the Limits of Pseudo Ground Truth in Visual Camera Re-Localisation
Eric Brachmann, Martin Humenberger, Carsten Rother, Torsten Sattler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6218-6228
Benchmark datasets that measure camera pose accuracy have driven progress in visual re-localisation research. To obtain poses for thousands of images, it is common to use a reference algorithm to generate pseudo ground truth. Popular choices include Structure-from-Motion (SfM) and Simultaneous-Localisation-and-Mapping (SLAM) using additional sensors like depth cameras if available. Re-localisation benchmarks thus measure how well each method replicates the results of the reference algorithm. This begs the question whether the choice of the reference algorithm favours a certain family of re-localisation methods. This paper analyzes two widely used re-localisation datasets and shows that evaluation outcomes indeed vary with the choice of the reference algorithm. We thus question common beliefs in the re-localisation literature, namely that learning-based scene coordinate regression outperforms classical feature-based methods, and that RGB-D- based methods outperform RGB-based methods. We argue that any claims on ranking re-localisation methods should take the type of the reference algorithm, and the similarity of the methods to the reference algorithm, into account.
****
An Elastica Geodesic Approach With Convexity Shape Prior
Da Chen, Laurent D. Cohen, Jean-Marie Mirebeau, Xue-Cheng Tai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6900-6909
The minimal geodesic models based on the Eikonal equations are capable of finding suitable solutions in various image segmentation scenarios. Existing geodesic-based segmentation approaches usually exploit the image features in conjunction with geometric regularization terms (such as curve length or elastica length) for computing geodesic paths. In this paper, we consider a more complicated problem: finding simple and closed geodesic curves which are imposed a convexity shape prior. The proposed approach relies on an orientation-lifting strategy, by which a planar curve can be mapped to an high-dimensional orientation space. The convexity shape prior serves as a constraint for the construction of local metrics. The geodesic curves in the lifted space then can be efficiently computed throug

h the fast marching method. In addition, we introduce a way to incorporate regio n-based homogeneity features into the proposed geodesic model so as to solve the region-based segmentation issues with shape prior constraints.

****

AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing S un, Qian Li, Errui Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6649-6658

Fast arbitrary neural style transfer has attracted widespread attention from aca demic, industrial and art communities due to its flexibility in enabling various applications. Existing solutions either attentively fuse deep style feature int o deep content feature without considering feature distributions, or adaptively normalize deep content feature according to the style such that their global sta tistic information is matched. Although effective, leaving shallow feature unexp lored or without locally considering feature statistics, they are prone to suffe r from unnatural output with unpleasing local distortions. To alleviate this pro blem, in this paper, we propose a novel Adaptive Attention Normalization (AdaAtt N) module to adaptively perform attentive normalization on per-point basis. Spec ifically, spatial attention score is learnt from both shallow and deep features of content and style images. Then per-point weighted statistics are calculated b y regarding a style feature point as a distribution of attention-weighted output of all style feature points. Finally, the content feature is normalized so that they demonstrate the same local feature statistics as the calculated per-point weighted style feature statistics. Besides, a novel local feature loss is derive d based on AdaAttN to enhance local visual quality. We also extend AdaAttN to be ready for video style transfer with slight modifications. Extensive experiments demonstrate that our method achieves state-of-the-art arbitrary image/video sty le transfer. Codes and models will be available.

****

PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recogni tion

Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D. Castillo, Rama Chellappa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15087-15096

Face recognition networks encode information about sensitive attributes while be ing trained for identity classification. Such encoding has two major issues: (a) it makes the face representations susceptible to privacy leakage (b) it appears to contribute to bias in face recognition. However, existing bias mitigation ap proaches generally require end-to-end training and are unable to achieve high ve rification accuracy. Therefore, we present a descriptor-based adversarial de-bia sing approach called `Protected Attribute Suppression System (PASS)'. PASS can b e trained on top of descriptors obtained from any previously trained high-perfor ming network to classify identities and simultaneously reduce encoding of sensit ive attributes. This eliminates the need for end-to-end training. As a component of PASS, we present a novel discriminator training strategy that discourages a network from encoding protected attribute information. We show the efficacy of P ASS to reduce gender and skintone information in descriptors from SOTA face reco gnition networks like Arcface. As a result, PASS descriptors outperform existing baselines in reducing gender and skintone bias on the IJB-C dataset, while main taining a high verification accuracy.

****

Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection

Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, Xu-Cheng Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1305 -1314

Arbitrary shape text detection is a challenging task due to the high complexity and variety of scene texts. In this work, we propose a novel adaptive boundary p roposal network for arbitrary shape text detection, which can learn to directly produce accurate boundary for arbitrary shape text without any post-processing. Our method mainly consists of a boundary proposal model and an innovative adapti

ve boundary deformation model. The boundary proposal model constructed by multi-layer dilated convolutions is adopted to produce prior information (including classification map, distance field, and direction field) and coarse boundary proposals. The adaptive boundary deformation model is an encoder-decoder network, in which the encoder mainly consists of a Graph Convolutional Network (GCN) and a Recurrent Neural Network (RNN). It aims to perform boundary deformation in an iterative way for obtaining text instance shape guided by prior information from the boundary proposal model. In this way, our method can directly and efficiently generate accurate text boundaries without complex post-processing. Extensive experiments on publicly available datasets demonstrate the state-of-the-art performance of our method.

****

Video Matting via Consistency-Regularized Graph Neural Networks
Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4902-4911

Learning temporally consistent foreground opacity from videos, i.e., video matting, has drawn great attention due to the blossoming of video conferencing. Previous approaches are built on top of image matting models, which fail in maintaining the temporal coherence when being adapted to videos. They either utilize the optical flow to smooth frame-wise prediction, where the performance is dependent on the selected optical flow model; or naively combine feature maps from multiple frames, which does not model well the correspondence of pixels in adjacent frames. In this paper, we propose to enhance the temporal coherence by Consistency-Regularized Graph Neural Networks (CRGNN) with the aid of a synthesized video matting dataset. CRGNN utilizes Graph Neural Networks (GNN) to relate adjacent frames such that pixels or regions that are incorrectly predicted in one frame can be corrected by leveraging information from its neighboring frames. To generalize our model from synthesized videos to real-world videos, we propose a consistency regularization technique to enforce the consistency on the alpha and foreground when blending them with different backgrounds. To evaluate the efficacy of CRGNN, we further collect a real-world dataset with annotated alpha mattes. Compared with state-of-the-art methods that require hand-crafted trimaps or backgrounds for modeling training, CRGNN generates favorably results with the help of unlabeled real training dataset.

****

Probabilistic Monocular 3D Human Pose Estimation With Normalizing Flows
Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, Bastian Wandt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11199-11208

3D human pose estimation from monocular images is a highly ill-posed problem due to depth ambiguities and occlusions. Nonetheless, most existing works ignore these ambiguities and only estimate a single solution. In contrast, we generate a diverse set of hypotheses that represents the full posterior distribution of feasible 3D poses. To this end, we propose a normalizing flow based method that exploits the deterministic 3D-to-2D mapping to solve the ambiguous inverse 2D-to-3D problem. Additionally, uncertain detections and occlusions are effectively modeled by incorporating uncertainty information of the 2D detector as condition. Further keys to success are a learned 3D pose prior and a generalization of the best-of-M loss. We evaluate our approach on the two benchmark datasets Human3.6M and MPI-INF-3DHP, outperforming all comparable methods in most metrics. The implementation is available on GitHub.

****

Inverting a Rolling Shutter Camera: Bring Rolling Shutter Images to High Framerate Global Shutter Video
Bin Fan, Yuchao Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4228-4237

Rolling shutter (RS) images can be viewed as the result of the row-wise combination of global shutter (GS) images captured by a virtual moving GS camera over the period of camera readout time. The RS effect brings tremendous difficulties fo

r the downstream applications. In this paper, we propose to invert the above RS imaging mechanism, i.e., recovering a high framerate GS video from consecutive RS images to achieve RS temporal super-resolution (RSSR). This extremely challenging problem, e.g., recovering 1440 GS images from two 720-height RS images, is far from being solved end-to-end. To address this challenge, we exploit the geometric constraint in the RS camera model, thus achieving geometry-aware inversion. Specifically, we make three contributions in resolving the above difficulties: (i) formulating the bidirectional RS undistortion flows under the constant velocity motion model, (ii) building the connection between the RS undistortion flow and optical flow via a scaling operation, and (iii) developing a mutual conversion scheme between varying RS undistortion flows that correspond to different scanlines. Building upon these formulations, we propose the first RS temporal super-resolution network in a cascaded structure to extract high framerate global shutter video. Our method explores the underlying spatio-temporal geometric relationships within a deep learning framework, where no extra supervision besides the middle-scanline ground truth GS image is needed. Essentially, our method can be very efficient for explicit propagation to generate GS images under any scanline. Experimental results on both synthetic and real data show that our method can produce high-quality GS image sequences with rich details, outperforming state-of-the-art methods.

****

Human Detection and Segmentation via Multi-View Consensus

Isinsu Katircioglu, Helge Rhodin, Jörg Spörri, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2855-2864

Self-supervised detection and segmentation of foreground objects aims for accuracy without annotated training data. However, existing approaches predominantly rely on restrictive assumptions on appearance and motion. For scenes with dynamic activities and camera motion, we propose a multi-camera framework in which geometric constraints are embedded in the form of multi-view consistency during training via coarse 3D localization in a voxel grid and fine-grained offset regression. In this manner, we learn a joint distribution of proposals over multiple views. At inference time, our method operates on single RGB images. We outperform state-of-the-art techniques both on images that visually depart from those of standard benchmarks and on those of the classical Human3.6M dataset.

****

GDP: Stabilized Neural Network Pruning via Gates With Differentiable Polarization

Yi Guo, Huan Yuan, Jianchao Tan, Zhangyang Wang, Sen Yang, Ji Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5239-5250

Model compression techniques are recently gaining explosive attention for obtaining efficient AI models for various real time applications. Channel pruning is one important compression strategy, and widely used in slimming various DNNs. Previous gate-based or importance-based pruning methods aim to remove channels whose "importance" are smallest. However, it remains unclear what criteria the channel importance should be measured on, leading to various channel selection heuristics. Some other sampling-based pruning methods deploy sampling strategy to train sub-nets, which often causes the training instability and the compressed model's degraded performance. In view of the research gaps, we present a new module named Gates with Differentiable Polarization (GDP), inspired by principled optimization ideas. GDP can be plugged before convolutional layers without bells and whistles, to control the on-and-off of each channel or whole layer block. During the training process, the polarization effect will drive a subset of gates to smoothly decrease to exactly zero, while other gates gradually stay away from zero by a large margin. When training terminates, those zero-gated channels can be painlessly removed, while other non-zero gates can be absorbed into the succeeding convolution kernel, causing completely no interruption to training nor damage to the trained model. Experiments conducted over CIFAR-10 and ImageNet datasets show that the proposed GDP algorithm achieves the state-of-the-art performance o

n various benchmark DNNs at a broad range of pruning ratios. We also apply GDP t
o DeepLabV3Plus-ResNet50 on the challenging Pascal VOC segmentation task, whose
test performance sees no drop (even slightly improved) with over 60% FLOPs savin
g.
****

From Two to One: A New Scene Text Recognizer With Visual Language Modeling Netwo
rk
Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, Yongdong Zhang
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2021, pp. 14194-14203
In this paper, we abandon the dominant complex language model and rethink the li
nguistic learning process in the scene text recognition. Different from previous
 methods considering the visual and linguistic information in two separate struc
tures, we propose a Visual Language Modeling Network (VisionLAN), which views th
e visual and linguistic information as a union by directly enduing the vision mo
del with language capability. Specially, we introduce the text recognition of ch
aracter-wise occluded feature maps in the training stage. Such operation guides
the vision model to use not only the visual texture of characters, but also the
linguistic information in visual context for recognition when the visual cues ar
e confused (e.g. occlusion, noise, etc.). As the linguistic information is acqui
red along with visual features without the need of extra language model, VisionL
AN significantly improves the speed by 39% and adaptively considers the linguist
ic information to enhance the visual features for accurate recognition. Furtherm
ore, an Occlusion Scene Text (OST) dataset is proposed to evaluate the performan
ce on the case of missing character-wise visual cues. The state-of-the-art resul
ts on several benchmarks prove our effectiveness. Code and dataset are available
 at https://github.com/wangyuxin87/VisionLAN.
****

GRF: Learning a General Radiance Field for 3D Representation and Rendering
Alex Trevithick, Bo Yang; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2021, pp. 15182-15192
We present a simple yet powerful neural network that implicitly represents and r
enders 3D objects and scenes only from 2D observations. The network models 3D ge
ometries as a general radiance field, which takes a set of 2D images with camera
 poses and intrinsics as input, constructs an internal representation for each p
oint of the 3D space, and then renders the corresponding appearance and geometry
 of that point viewed from an arbitrary position. The key to our approach is to
learn local features for each pixel in 2D images and to then project these featu
res to 3D points, thus yielding general and rich point representations. We addit
ionally integrate an attention mechanism to aggregate pixel features from multip
le 2D views, such that visual occlusions are implicitly taken into account. Exte
nsive experiments demonstrate that our method can generate high-quality and real
istic novel views for novel objects, unseen categories and challenging real-worl
d scenes.
****

Neural Strokes: Stylized Line Drawing of 3D Shapes
Difan Liu, Matthew Fisher, Aaron Hertzmann, Evangelos Kalogerakis; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 142
04-14213
This paper introduces a model for producing stylized line drawings from 3D shape
s. The model takes a 3D shape and a viewpoint as input, and outputs a drawing wi
th textured strokes, with variations in stroke thickness, deformation, and color
 learned from an artist's style. The model is fully differentiable. We train its
 parameters from a single training drawing of another 3D shape. We show that, in
 contrast to previous image-based methods, the use of a geometric representation
 of 3D shape and 2D strokes allows the model to transfer important aspects of sh
ape and texture style while preserving contours. Our method outputs the resultin
g drawing in a vector representation, enabling richer downstream analysis or edi
ting in interactive applications.
****

Multimodal Knowledge Expansion
Zihui Xue, Sucheng Ren, Zhengqi Gao, Hang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 854-863
The popularity of multimodal sensors and the accessibility of the Internet have brought us a massive amount of unlabeled multimodal data. Since existing datasets and well-trained models are primarily unimodal, the modality gap between a unimodal network and unlabeled multimodal data poses an interesting problem: how to transfer a pre-trained unimodal network to perform the same task on unlabeled multimodal data? In this work, we propose multimodal knowledge expansion (MKE), a knowledge distillation-based framework to effectively utilize multimodal data without requiring labels. Opposite to traditional knowledge distillation, where the student is designed to be lightweight and inferior to the teacher, we observe that the multimodal student model consistently rectifies pseudo labels and generalizes better than its teacher. Extensive experiments on four tasks and different modalities verify this finding. Furthermore, we connect the mechanism of MKE to semi-supervised learning and offer both empirical and theoretical explanations to understand the expansion capability of a multimodal student.
****

Learning To Bundle-Adjust: A Graph Network Approach to Faster Optimization of Bundle Adjustment for Vehicular SLAM
Tetsuya Tanaka, Yukihiro Sasagawa, Takayuki Okatani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6250-6259
Bundle adjustment (BA) occupies a large portion of SfM and visual SLAM's total execution time. Local BA over the latest several keyframes plays a crucial role in visual SLAM. Its execution time should be sufficiently short for robust tracking; this is especially critical for embedded systems with a limited computational resource. This study proposes a learning-based method using a graph network that can replace conventional optimization-based BA and works faster. The graph network operates on a graph consisting of the nodes of keyframes and landmarks and the edges of the latter's visibility from the former. The graph network receives the parameters' initial values as inputs and predicts the updates to their optimal values. We design an intermediate representation of inputs inspired by the normal equation of the Levenberg-Marquardt method. We use the sum of reprojection errors as a loss function to train the graph network. The experiments show that the proposed method outputs parameter estimates with slightly inferior accuracy in 1/60-1/10 of time compared with the conventional BA.
****

MosaicOS: A Simple and Effective Use of Object-Centric Images for Long-Tailed Object Detection
Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, Wei-Lun Chao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 417-427
Many objects do not appear frequently enough in complex scenes (e.g., certain handbags in living rooms) for training an accurate object detector, but are often found frequently by themselves (e.g., in product images). Yet, these object-centric images are not effectively leveraged for improving object detection in scene-centric images. In this paper, we propose Mosaic of Object-centric images as Scene-centric images (MosaicOS), a simple and novel framework that is surprisingly effective at tackling the challenges of long-tailed object detection. Keys to our approach are three-fold: (i) pseudo scene-centric image construction from object-centric images for mitigating domain differences, (ii) high-quality bounding box imputation using the object-centric images' class labels, and (iii) a multi-stage training procedure. On LVIS object detection (and instance segmentation), MosaicOS leads to a massive 60% (and 23%) relative improvement in average precision for rare object categories. We also show that our framework can be compatibly used with other existing approaches to achieve even further gains. Our pre-trained models are publicly available at https://github.com/czhang0528/MosaicOS/.
****

Bringing Events Into Video Deblurring With Non-Consecutively Blurry Frames
Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S. Ren, Ping Luo, Wangmeng Zuo; Proc

eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4531-4540
Recently, video deblurring has attracted considerable research attention, and several works suggest that events at high time rate can benefit deblurring. In this paper, we develop a principled framework D2Nets for video deblurring to exploit non-consecutively blurry frames, and propose a flexible event fusion module (EFM) to bridge the gap between event-driven and video deblurring. In D2Nets, we propose to first detect nearest sharp frames (NSFs) using a bidirectional LSTM detector, and then perform deblurring guided by NSFs. Furthermore, the proposed EFM is flexible to be incorporated into D2Nets, in which events can be leveraged to notably boost the deblurring performance. EFM can also be easily incorporated into existing deblurring networks, making event-driven deblurring task benefit from state-of-the-art deblurring methods. On synthetic and real-world blurry datasets, our methods achieve better results than competing methods, and EFM not only benefits D2Nets but also significantly improves the competing deblurring networks.
****

SPG: Unsupervised Domain Adaptation for 3D Object Detection via Semantic Point Generation
Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R. Qi, Dragomir Anguelov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15446-15456
In autonomous driving, a LiDAR-based object detector should perform reliably at different geographic locations and under various weather conditions. While recent 3D detection research focuses on improving performance within a single domain, our study reveals that the performance of modern detectors can drop drastically cross-domain. In this paper, we investigate unsupervised domain adaptation (UDA) for LiDAR-based 3D object detection. On the Waymo Domain Adaptation dataset, we identify the deteriorating point cloud quality as the root cause of the performance drop. To address this issue, we present Semantic Point Generation (SPG), a general approach to enhance the reliability of LiDAR detectors against domain shifts. Specifically, SPG generates semantic points at the predicted foreground regions and faithfully recovers missing parts of the foreground objects, which are caused by phenomena such as occlusions, low reflectance, or weather interference. By merging the semantic points with the original points, we obtain an augmented point cloud, which can be directly consumed by modern LiDAR-based detectors.
To validate the wide applicability of SPG, we experiment with two representative detectors, PointPillars and PV-RCNN. On the UDA task, SPG significantly improves both detectors across all object categories of interest and at all difficulty levels. SPG can also benefit object detection in the original domain. On the Waymo Open Dataset and KITTI, SPG improves 3D detection results of these two methods across all categories. Combined with PV-RCNN, SPG achieves state-of-the-art 3D detection results on KITTI.
****

Extreme-Quality Computational Imaging via Degradation Framework
Shiqi Chen, Huajun Feng, Keming Gao, Zhihai Xu, Yueting Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2632-2641
To meet the space limitation of optical elements, free-form surfaces or high-order aspherical lenses are adopted in mobile cameras to compress volume. However, the application of free-form surfaces also introduces the problem of image quality mutation. Existing model-based deconvolution methods are inefficient in dealing with the degradation that shows a wide range of spatial variants over regions. And the deep learning techniques in low-level and physics-based vision suffer from a lack of accurate data. To address this issue, we develop a degradation framework to estimate the spatially variant point spread functions (PSFs) of mobile cameras. When input extreme-quality digital images, the proposed framework generates degraded images sharing a common domain with real-world photographs. Supplied with the synthetic image pairs, we design a Field-Of-View shared kernel prediction network (FOV-KPN) to perform spatial-adaptive reconstruction on real deg

raded photos. Extensive experiments demonstrate that the proposed approach achie
ves extreme-quality computational imaging and outperforms the state-of-the-art m
ethods. Furthermore, we illustrate that our technique can be integrated into exi
sting postprocessing systems, resulting in significantly improved visual quality
.
****

Direct Differentiable Augmentation Search
Aoming Liu, Zehao Huang, Zhiwu Huang, Naiyan Wang; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 12219-12228
Data augmentation has been an indispensable tool to improve the performance of d
eep neural networks, however the augmentation can hardly transfer among differen
t tasks and datasets. Consequently, a recent trend is to adopt AutoML technique
to learn proper augmentation policy without extensive hand-crafted tuning. In th
is paper, we propose an efficient differentiable search algorithm called Direct
Differentiable Augmentation Search (DDAS). It utilizes meta-learning with one-st
ep gradient update and continuous relaxation to the expected training loss for e
fficient search. Our DDAS could achieve efficient augmentation search without ap
proximations such as Gumbel-Softmax or second order gradient approximation. To f
urther reduce the adverse effect of improper augmentations, we organize the sear
ch space into a two level hierarchy, in which we first decide whether to apply a
ugmentation, and then determine the specific augmentation policy. On standard im
age classification benchmarks, our DDAS achieves state-of-the-art performance an
d efficiency tradeoff while reducing the search cost dramatically, e.g. 0.15 GPU
 hours for CIFAR-10. In addition, we also use DDAS to search augmentation for ob
ject detection task and achieve comparable performance with AutoAugment, while b
eing 1000x faster. Code will be released in https://github.com/zxcvfd13502/DDAS_
code.
****

The Functional Correspondence Problem
Zihang Lai, Senthil Purushwalkam, Abhinav Gupta; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2021, pp. 15772-15781
The ability to find correspondences in visual data is the essence of most comput
er vision tasks. But what are the right correspondences? The task of visual corr
espondence is well defined for two different images of same object instance. In
case of two images of objects belonging to same category, visual correspondence
is reasonably well-defined in most cases. But what about correspondence between
two objects of completely different category -- e.g., a shoe and a bottle? Does
there exist any correspondence? Inspired by humans' ability to: (a) generalize b
eyond semantic categories and; (b) infer functional affordances, we introduce th
e problem of functional correspondences in this paper. Given images of two objec
ts, we ask a simple question: what is the set of correspondences between these t
wo images for a given task? For example, what are the correspondences between a
bottle and shoe for the task of pounding or the task of pouring. We introduce a
new dataset: FunKPoint that has ground truth correspondences for 10 tasks and 20
 object categories. We also introduce a modular task-driven representation for a
ttacking this problem and demonstrate that our learned representation is effecti
ve for this task. But most importantly, because our supervision signal is not bo
und by semantics, we show that our learned representation can generalize better
on few-shot classification problem. We hope this paper will inspire our communit
y to think beyond semantics and focus more on cross-category generalization and
learning representations for robotics tasks.
****

Detection and Continual Learning of Novel Face Presentation Attacks
Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, Wael Abd-Alma
geed; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2021, pp. 14851-14860
Advances in deep learning, combined with availability of large datasets, have le
d to impressive improvements in face presentation attack detection research. How
ever, state of the art face antispoofing systems are still vulnerable to novel t
ypes of attacks that are never seen during training. Moreover, even if such atta

cks are correctly detected, these systems lack the ability to adapt to newly enc
ountered attacks. The post-training ability of continually detecting new types o
f attacks and self-adaptation to identify these attack types, after the initial
detection phase, is highly appealing. In this paper, we enable a deep neural net
work to detect anomalies in the observed input data points as potential new type
s of attacks by suppressing the confidence-level of the network outside the trai
ning samples' distribution. We then use experience replay to update the model to
 incorporate knowledge about new types of attacks without forgetting the past le
arned attack types. Experimental results are provided to demonstrate the effecti
veness of the proposed method on the OULU and Idiap datasets as well as a newly
introduced dataset, all of which exhibit a variety of attack types.
****

Adaptive Adversarial Network for Source-Free Domain Adaptation
Haifeng Xia, Handong Zhao, Zhengming Ding; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 9010-9019
Unsupervised Domain Adaptation solves knowledge transfer along with the coexiste
nce of well-annotated source domain and unlabeled target instances. However, the
 source domain in many practical applications is not always accessible due to da
ta privacy or the insufficient memory storage for small devices. This scenario d
efined as Source-free Domain Adaptation only allows accessing the well-trained s
ource model for target learning. To address the challenge of source data unavail
ability, we develop an Adaptive Adversarial Network (A2Net) including three comp
onents. Specifically, the first one named Adaptive Adversarial Inference seeks a
 target-specific classifier to advance the recognition of samples which the prov
ided source-specific classifier difficultly identifies. Then, the Contrastive Ca
tegory-wise Matching module exploits the positive relation of every two target i
mages to enforce the compactness of subspace for each category. Thirdly, Self-Su
pervised Rotation facilitates the model to learn additional semantics from targe
t images by themselves. Extensive experiments on the popular cross-domain benchm
arks verify the effectiveness of our proposed model on solving adaptation task w
ithout any source data.
****

Painting From Part
Dongsheng Guo, Haoru Zhao, Yunhao Cheng, Haiyong Zheng, Zhaorui Gu, Bing Zheng;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2021, pp. 14779-14788
This paper studies the problem of painting the whole image from part of it, name
ly painting from part or part-painting for short, involving both inpainting and
outpainting. To address the challenge of taking full advantage of both informati
on from local domain (part) and knowledge from global domain (dataset), we propo
se a novel part-painting method according to the observations of relationship be
tween part and whole, which consists of three stages: part-noise restarting, par
t-feature repainting, and part-patch refining, to paint the whole image by lever
aging both feature-level and patch-level part as well as powerful representation
 ability of generative adversarial network. Extensive ablation studies show effi
cacy of each stage, and our method achieves state-of-the-art performance on both
 inpainting and outpainting benchmarks with free-form parts, including our new m
ask dataset for irregular outpainting. Our code and dataset are available at htt
ps://github.com/zhenglab/partpainting.
****

Attack-Guided Perceptual Data Generation for Real-World Re-Identification
Yukun Huang, Xueyang Fu, Zheng-Jun Zha; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2021, pp. 215-224
In unconstrained real-world surveillance scenarios, person re-identification (Re
-ID) models usually suffer from different low-level perceptual variations, e.g.,
 cross-resolution and insufficient lighting. Due to the limited variation range
of training data, existing models are difficult to generalize to scenes with unk
nown perceptual interference types. To address the above problem, in this paper,
 we propose two disjoint data-generation ways to complement existing training sa
mples to improve the robustness of Re-ID models. Firstly, considering the sparsi

ty and imbalance of samples in the perceptual space, a dense resampling method f
rom the estimated perceptual distribution is performed. Secondly, to dig more re
presentative generated samples for identity representation learning, we introduc
e a graph-based white-box attacker to guide the data generation process with int
ra-batch ranking and discriminate attention. In addition, two synthetic-to-real
feature constraints are introduced into the Re-ID training to prevent the genera
ted data from bringing domain bias. Our method is effective, easy-to-implement,
and independent of the specific network architecture. Applying our approach to a
 ResNet-50 baseline can already achieve competitive results, surpassing state-of
-the-art methods by +1.2% at Rank-1 on the MLR-CUHK03 dataset.
****

Parallel Multi-Resolution Fusion Network for Image Inpainting
Wentao Wang, Jianfu Zhang, Li Niu, Haoyu Ling, Xue Yang, Liqing Zhang; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
 14559-14568
Conventional deep image inpainting methods are based on auto-encoder architectur
e, in which the spatial details of images will be lost in the down-sampling proc
ess, leading to the degradation of generated results. Also, the structure inform
ation in deep layers and texture information in shallow layers of the auto-encod
er architecture can not be well integrated. Differing from the conventional imag
e inpainting architecture, we design a parallel multi-resolution inpainting netw
ork with multi-resolution partial convolution, in which low-resolution branches
focus on the global structure while high-resolution branches focus on the local
texture details. All these high- and low-resolution streams are in parallel and
fused repeatedly with multi-resolution masked representation fusion so that the
reconstructed images are semantically robust and textually plausible. Experiment
al results show that our method can effectively fuse structure and texture infor
mation, producing more realistic results than state-of-the-art methods.
****

Joint Topology-Preserving and Feature-Refinement Network for Curvilinear Structu
re Segmentation
Mingfei Cheng, Kaili Zhao, Xuhong Guo, Yajing Xu, Jun Guo; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7147-7156
Curvilinear structure segmentation (CSS) is under semantic segmentation, whose a
pplications include crack detection, aerial road extraction, and biomedical imag
e segmentation. In general, geometric topology and pixel-wise features are two c
ritical aspects of CSS. However, most semantic segmentation methods only focus o
n enhancing feature representations while existing CSS techniques emphasize pres
erving topology alone. In this paper, we present a Joint Topology-preserving and
 Feature-refinement Network (JTFN) that jointly models global topology and refin
ed features based on an iterative feedback learning strategy. Specifically, we e
xplore the structure of objects to help preserve corresponding topologies of pre
dicted masks, thus design a reciprocative two-stream module for CSS and boundary
 detection. In addition, we introduce such topology-aware predictions as feedbac
k guidance that refines attentive features by supplementing and enhancing salien
cies. To the best of our knowledge, this is the first work that jointly addresse
s topology preserving and feature refinement for CSS. We evaluate JTFN on four d
atasets of diverse applications: Crack500, CrackTree200, Roads, and DRIVE. Resul
ts show that JTFN performs best in comparison with alternative methods. Code is
available.
****

MT-ORL: Multi-Task Occlusion Relationship Learning
Panhe Feng, Qi She, Lei Zhu, Jiaxin Li, Lin Zhang, Zijian Feng, Changhu Wang, Ch
unpeng Li, Xuejing Kang, Anlong Ming; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2021, pp. 9364-9373
Retrieving occlusion relation among objects in a single image is challenging due
 to sparsity of boundaries in image. We observe two key issues in existing works
: firstly, lack of an architecture which can exploit the limited amount of coupl
ing in the decoder stage between the two subtasks, namely occlusion boundary ext
raction and occlusion orientation prediction, and secondly, improper representat

ion of occlusion orientation. In this paper, we propose a novel architecture called Occlusion-shared and Path-separated Network (OPNet), which solves the first issue by exploiting rich occlusion cues in shared high-level features and structured spatial information in task-specific low-level features. We then design a simple but effective orthogonal occlusion representation (OOR) to tackle the second issue. Our method surpasses the state-of-the-art methods by 6.1%/8.3% Boundary-AP and 6.5%/10% Orientation-AP on standard PIOD/BSDS ownership datasets. Code is available at https://github.com/fengpanhe/MT-ORL.

****

Weakly Supervised Human-Object Interaction Detection in Video via Contrastive Spatiotemporal Regions

Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, Bryan Russell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1845-1855

We introduce the task of weakly supervised learning for detecting human and object interactions in videos. Our task poses unique challenges as a system does not know what types of human-object interactions are present in a video or the actual spatiotemporal location of the human and object. To address these challenges, we introduce a contrastive weakly supervised training loss that aims to jointly associate spatiotemporal regions in a video with an action and object vocabulary and encourage temporal continuity of the visual appearance of moving objects as a form of self-supervision. To train our model, we introduce a dataset comprising over 6.5k videos with human-object interaction annotations that have been semi-automatically curated from sentence captions associated with the videos. We demonstrate improved performance over weakly supervised baselines adapted to our task on our video dataset.

****

Generative Layout Modeling Using Constraint Graphs

Wamiq Para, Paul Guerrero, Tom Kelly, Leonidas J. Guibas, Peter Wonka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6690-6700

We propose a new generative model for layout generation. We generate layouts in three steps. First, we generate the layout elements as nodes in a layout graph. Second, we compute constraints between layout elements as edges in the layout graph. Third, we solve for the final layout using constrained optimization. For the first two steps, we build on recent transformer architectures. The layout optimization implements the constraints efficiently. We show three practical contributions compared to the state of the art: our work requires no user input, produces higher quality layouts, and enables many novel capabilities for conditional layout generation.

****

ZFlow: Gated Appearance Flow-Based Virtual Try-On With 3D Priors

Ayush Chopra, Rishabh Jain, Mayur Hemani, Balaji Krishnamurthy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5433-5442

Image-based virtual try-on involves synthesizing perceptually convincing images of a model wearing a particular garment and has garnered significant research interest due to its immense practical applicability. Recent methods involve a two-stage process: i) warping of the garment to align with the model ii) texture fusion of the warped garment and target model to generate the try-on output. Issues arise due to the non-rigid nature of garments and the lack of geometric information about the model or the garment. It often results in improper rendering of granular details. We propose ZFlow, an end-to-end framework, which seeks to alleviate these concerns regarding geometric and textural integrity (such as pose, depth-ordering, skin and neckline reproduction) through a combination of gated aggregation of hierarchical flow estimates termed Gated Appearance Flow, and dense structural priors at various stage of the network. ZFlow achieves state-of-the-art results as observed qualitatively, and on benchmark image quality measures (PSNR, SSIM, and FID scores). The paper also presents extensive comparisons with existing state-of-the-art including a detailed user study and ablation studies to

gauge the effectiveness of each of our contributions on multiple datasets
****
Overfitting the Data: Compact Neural Video Delivery via Content-Aware Feature Modulation
Jiaming Liu, Ming Lu, Kaixin Chen, Xiaoqi Li, Shizun Wang, Zhaoqing Wang, Enhua Wu, Yurong Chen, Chuang Zhang, Ming Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4631-4640
Internet video delivery has undergone a tremendous explosion of growth over the past few years. However, the quality of video delivery system greatly depends on the Internet bandwidth. Deep Neural Networks (DNNs) are utilized to improve the quality of video delivery recently. These methods divide a video into chunks, and stream LR video chunks and corresponding content-aware models to the client. The client runs the inference of models to super-resolve the LR chunks. Consequently, a large number of models are streamed in order to deliver a video. In this paper, we first carefully study the relation between models of different chunks, then we tactfully design a joint training framework along with the Content-aware Feature Modulation (CaFM) layer to compress these models for neural video delivery. With our method, each video chunk only requires less than 1% of original parameters to be streamed, achieving even better SR performance. We conduct extensive experiments across various SR backbones, video time length, and scaling factors to demonstrate the advantages of our method. Besides, our method can be also viewed as a new approach of video coding. Our primary experiments achieve better video quality compared with the commercial H.264 and H.265 standard under the same storage cost, showing the great potential of the proposed method. Code is available at: https://github.com/Neural-video-delivery/CaFM-Pytorch-ICCV2021
****
Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation
Weiyao Wang, Matt Feiszli, Heng Wang, Du Tran; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10776-10785
Current state-of-the-art object detection and segmentation methods work well under the closed-world assumption. This closed-world setting assumes that the list of object categories is available during training and deployment. However, many real-world applications require detecting or segmenting novel objects, i.e., object categories never seen during training. In this paper, we present, UVO (Unidentified Video Objects), a new benchmark for open-world class-agnostic object segmentation in videos. Besides shifting the focus to the open-world setup, UVO is significantly larger, providing approximately 6 times more videos compared with DAVIS, and 7 times more mask (instance) annotations per video compared with YouTube-VO(I)S. UVO is also more challenging as it includes many videos with crowded scenes and complex background motions. We also demonstrated that UVO can be used for other applications, such as object tracking and super-voxel segmentation. We believe that UVO is a versatile testbed for researchers to develop novel approaches for open-world class-agnostic object segmentation, and inspires new research directions towards a more comprehensive video understanding beyond classification and detection.
****
Weakly Supervised Relative Spatial Reasoning for Visual Question Answering
Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, Chitta Baral; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1908-1918
Vision-and-language (V&L) reasoning necessitates perception of visual concepts such as objects and actions, understanding semantics and language grounding, and reasoning about the interplay between the two modalities. One crucial aspect of visual reasoning is spatial understanding, which involves understanding relative locations of objects, i.e. implicitly learning the geometry of the scene. In this work, we evaluate the faithfulness of V&L models to such geometric understanding, by formulating the prediction of pair-wise relative locations of objects as a classification as well as a regression task. Our findings suggest that state-of-the-art transformer-based V&L models lack sufficient abilities to excel at this task. Motivated by this, we design two objectives as proxies for 3D spatial reasoning (SR) -- object centroid estimation, and relative position estimation, a

nd train V&L with weak supervision from off-the-shelf depth estimators. This lea ds to considerable improvements in accuracy for the "GQA" visual question answer ing challenge (in fully supervised, few-shot, and O.O.D settings) as well as imp rovements in relative spatial reasoning. Code and data will be released here.
****

Task-Aware Part Mining Network for Few-Shot Learning
Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, Feng Wu; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2021, pp. 8433-8442
Few-Shot Learning (FSL) aims at classifying samples into new unseen classes with only a handful of labeled samples available. However, most of the existing meth ods are based on the image-level pooled representation, yet ignore considerable local clues that are transferable across tasks. To address this issue, we propos e an end-to-end Task-aware Part Mining Network (TPMN) by integrating an automati c part mining process into the metric-based model for FSL. The proposed TPMN mod el enjoys several merits. First, we design a meta filter learner to generate tas k-aware part filters based on the task embedding in a meta-learning way. The tas k-aware part filters can adapt to any individual task and automatically mine tas k-related local parts even for an unseen task. Second, an adaptive importance ge nerator is proposed to identify key local parts and assign adaptive importance w eights to different parts. To the best of our knowledge, this is the first work to automatically exploit the task-aware local parts in a meta-learning way for F SL. Extensive experimental results on four standard benchmarks demonstrate that the proposed model performs favorably against state-of-the-art FSL methods.
****

Cascade Image Matting With Deformable Graph Refinement
Zijian Yu, Xuhui Li, Huijuan Huang, Wen Zheng, Li Chen; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2021, pp. 7167-7176
Image matting refers to the estimation of the opacity of foreground objects. It requires correct contours and fine details of foreground objects for the matting results. To better accomplish human image matting tasks, we propose the Cascade Image Matting Network with Deformable Graph Refinement(CasDGR), which can autom atically predict precise alpha mattes from single human images without any addit ional inputs. We adopt a network cascade architecture to perform matting from lo w-to-high resolution, which corresponds to coarse-to-fine optimization. We also introduce the Deformable Graph Refinement (DGR) module based on graph neural net works (GNNs) to overcome the limitations of convolutional neural networks (CNNs) . The DGR module can effectively capture long-range relations and obtain more gl obal and local information to help produce finer alpha mattes. We also reduce th e computation complexity of the DGR module by dynamically predicting the neighbo rs and apply DGR module to higher-resolution features. Experimental results demo nstrate the ability of our CasDGR to achieve state-of-the-art performance on syn thetic datasets and produce good results on real human images.
****

Geometric Unsupervised Domain Adaptation for Semantic Segmentation
Vitor Guizilini, Jie Li, Rare■ Ambru■, Adrien Gaidon; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 8537-8547
Simulators can efficiently generate large amounts of labeled synthetic data with perfect supervision for hard-to-label tasks like semantic segmentation. However , they introduce a domain gap that severely hurts real-world performance. We pro pose to use self-supervised monocular depth estimation as a proxy task to bridge this gap and improve sim-to-real unsupervised domain adaptation (UDA). Our Geom etric Unsupervised Domain Adaptation method (GUDA) learns a domain-invariant rep resentation via a multi-task objective combining synthetic semantic supervision with real-world geometric constraints on videos. GUDA establishes a new state of the art in UDA for semantic segmentation on three benchmarks, outperforming met hods that use domain adversarial learning, self-training, or other self-supervis ed proxy tasks. Furthermore, we show that our method scales well with the qualit y and quantity of synthetic data while also improving depth prediction.
****

A Hierarchical Variational Neural Uncertainty Model for Stochastic Video Predict

ion
Moitreya Chatterjee, Narendra Ahuja, Anoop Cherian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9751-9761
Predicting the future frames of a video is a challenging task, in part due to the underlying stochastic real-world phenomena. Prior approaches to solve this task typically estimate a latent prior characterizing this stochasticity, however do not account for the predictive uncertainty of the (deep learning) model. Such approaches often derive the training signal from the mean-squared error (MSE) between the generated frame and the ground truth, which can lead to sub-optimal training, especially when the predictive uncertainty is high. Towards this end, we introduce Neural Uncertainty Quantifier (NUQ) - a stochastic quantification of the model's predictive uncertainty, and use it to weigh the MSE loss. We propose a hierarchical, variational framework to derive NUQ in a principled manner using a deep, Bayesian graphical model. Our experiments on three benchmark stochastic video prediction datasets show that our proposed framework trains more effectively compared to the state-of-the-art models (especially when the training sets are small), while demonstrating better video generation quality and diversity against several evaluation metrics.
****
PX-NET: Simple and Efficient Pixel-Wise Training of Photometric Stereo Networks
Fotios Logothetis, Ignas Budvytis, Roberto Mecca, Roberto Cipolla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12757-12766
Retrieving accurate 3D reconstructions of objects from the way they reflect light is a very challenging task in computer vision. Despite more than four decades since the definition of the Photometric Stereo problem, most of the literature has had limited success when global illumination effects such as cast shadows, self-reflections and ambient light come into play, especially for specular surfaces. Recent approaches have leveraged the capabilities of deep learning in conjunction with computer graphics in order to cope with the need of a vast number of training data to invert the image irradiance equation and retrieve the geometry of the object. However, rendering global illumination effects is a slow process which can limit the amount of training data that can be generated. In this work we propose a novel pixel-wise training procedure for normal prediction by replacing the training data (observation maps) of globally rendered images with independent per-pixel generated data. We show that global physical effects can be approximated on the observation map domain and this simplifies and speeds up the data creation procedure. Our network, PX-NET, achieves state-of-the-art performance compared to other pixelwise methods on synthetic datasets, as well as the DiLiGenT real dataset on both dense and sparse light settings.
****
Dynamic Surface Function Networks for Clothed Human Bodies
Andrei Burov, Matthias Nießner, Justus Thies; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10754-10764
We present a novel method for temporal coherent reconstruction and tracking of clothed humans. Given a monocular RGB-D sequence, we learn a person-specific body model which is based on a dynamic surface function network. To this end, we explicitly model the surface of the person using a multi-layer perceptron (MLP) which is embedded into the canonical space of the SMPL body model. With classical forward rendering, the represented surface can be rasterized using the topology of a template mesh. For each surface point of the template mesh, the MLP is evaluated to predict the actual surface location. To handle pose-dependent deformations, the MLP is conditioned on the SMPL pose parameters. We show that this surface representation as well as the pose parameters can be learned in a self-supervised fashion using the principle of analysis-by-synthesis and differentiable rasterization. As a result, we are able to reconstruct a temporally coherent mesh sequence from the input data. The underlying surface representation can be used to synthesize new animations of the reconstructed person including pose-dependent deformations.
****

Preservational Learning Improves Self-Supervised Medical Image Models by Reconstructing Diverse Contexts
Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, Yizhou Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3499-3509
Preserving maximal information is the basic principle of designing self-supervised learning methodologies. To reach this goal, contrastive learning adopts an implicit way which is contrasting image pairs. However, we believe it is not fully optimal to simply use the contrastive estimation for preservation. Moreover, it is necessary and complemental to introduce an explicit solution to preserve more information. From this perspective, we introduce Preservational Learning to reconstruct diverse image contexts in order to preserve more information in learned representations. Together with the contrastive loss, we present Preservational Contrastive Representation Learning (PCRL) for learning self-supervised medical representations. PCRL provides very competitive results under the pretraining-finetuning protocol, outperforming both self-supervised and supervised counterparts in 5 classification/segmentation tasks substantially.
****

Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework
Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Yang Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3365-3374
Localizing individuals in crowds is more in accordance with the practical demands of subsequent high-level crowd analysis tasks than simply counting. However, existing localization based methods relying on intermediate representations (i.e., density maps or pseudo boxes) serving as learning targets are counter-intuitive and error-prone. In this paper, we propose a purely point-based framework for joint crowd counting and individual localization. For this framework, instead of merely reporting the absolute counting error at image level, we propose a new metric, called density Normalized Average Precision (nAP), to provide more comprehensive and more precise performance evaluation. Moreover, we design an intuitive solution under this framework, which is called Point to Point Network (P2PNet). P2PNet discards superfluous steps and directly predicts a set of point proposals to represent heads in an image, being consistent with the human annotation results. By thorough analysis, we reveal the key step towards implementing such a novel idea is to assign optimal learning targets for these proposals. Therefore, we propose to conduct this crucial association in an one-to-one matching manner using the Hungarian algorithm. The P2PNet not only significantly surpasses state-of-the-art methods on popular counting benchmarks, but also achieves promising localization accuracy. The codes will be available at: https://github.com/TencentYoutuResearch/CrowdCounting-P2PNet.
****

The Right To Talk: An Audio-Visual Transformer Approach
Thanh-Dat Truong, Chi Nhan Duong, The De Vu, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1105-1114
Turn-taking has played an essential role in structuring the regulation of a conversation. The task of identifying the main speaker (who is properly taking his/her turn of speaking) and the interrupters (who are interrupting or reacting to the main speaker's utterances) remains a challenging task. Although some prior methods have partially addressed this task, there still remain some limitations. Firstly, a direct association of Audio and Visual features may limit the correlations to be extracted due to different modalities. Secondly, the relationship across temporal segments helping to maintain the consistency of localization, separation and conversation contexts is not effectively exploited. Finally, the interactions between speakers that usually contain the tracking and anticipatory decisions about transition to a new speaker is usually ignored. Therefore, this work introduces a new Audio-Visual Transformer approach to the problem of localization and highlighting the main speaker in both audio and visual channels of a multi-speaker conversation video in the wild. The proposed method exploits different

types of correlations presented in both visual and audio signals. The temporal audio-visual relationships across spatial-temporal space are anticipated and optimized via the self-attention mechanism in a Transformer structure. Moreover, a newly collected dataset is introduced for the main speaker detection. To the best of our knowledge, it is one of the first studies that is able to automatically localize and highlight the main speaker in both visual and audio channels in multi-speaker conversation videos.

****

Neural Image Compression via Attentional Multi-Scale Back Projection and Frequency Decomposition

Ge Gao, Pei You, Rong Pan, Shunyuan Han, Yuanyuan Zhang, Yuchao Dai, Hojae Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14677-14686

In recent years, neural image compression emerges as a rapidly developing topic in computer vision, where the state-of-the-art approaches now exhibit superior compression performance than their conventional counterparts. Despite the great progress, current methods still have limitations in preserving fine spatial details for optimal reconstruction, especially at low compression rates. We make three contributions in tackling this issue. First, we develop a novel back projection method with attentional and multi-scale feature fusion for augmented representation power. Our back projection method recalibrates the current estimation by establishing feedback connections between high-level and low-level attributes in an attentional and discriminative manner. Second, we propose to decompose the input image and separately process the distinct frequency components, whose derived latents are recombined using a novel dual attention module, so that details inside regions of interest could be explicitly manipulated. Third, we propose a novel training scheme for reducing the latent rounding residual. Experimental results show that, when measured in PSNR, our model reduces BD-rate by 9.88% and 10.32% over the state-of-the-art method, and 4.12% and 4.32% over the latest coding standard Versatile Video Coding (VVC) on the Kodak and CLIC2020 Professional Validation dataset, respectively. Our approach also produces more visually pleasant images when optimized for MS-SSIM. The significant improvement upon existing methods shows the effectiveness of our method in preserving and remedying spatial information for enhanced compression quality.

****

Unpaired Learning for High Dynamic Range Image Tone Mapping

Yael Vinker, Inbar Huberman-Spiegelglas, Raanan Fattal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14657-14666

High dynamic range (HDR) photography is becoming increasingly popular and available by DSLR and mobile-phone cameras. While deep neural networks (DNN) have greatly impacted other domains of image manipulation, their use for HDR tone-mapping is limited due to the lack of a definite notion of ground-truth solution, which is needed for producing training data. In this paper we describe a new tone-mapping approach guided by the distinct goal of producing low dynamic range (LDR) renditions that best reproduce the visual characteristics of native LDR images. This goal enables the use of an unpaired adversarial training based on unrelated sets of HDR and LDR images, both of which are widely available and easy to acquire. In order to achieve an effective training under this minimal requirements, we introduce the following new steps and components: (i) a range-normalizing pre-process which estimates and applies a different level of curve-based compression, (ii) a loss that preserves the input content while allowing the network to achieve its goal, and (iii) the use of a more concise discriminator network, designed to promote the reproduction of low-level attributes native LDR possess. Evaluation of the resulting network demonstrates its ability to produce photo-realistic artifact-free tone-mapped images, and state-of-the-art performance on different image fidelity indices and visual distances.

****

Unsupervised Real-World Super-Resolution: A Domain Adaptation Perspective

Wei Wang, Haochen Zhang, Zehuan Yuan, Changhu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4318-4327

Most existing convolution neural network (CNN) based super-resolution (SR) metho
ds generate their paired training dataset by artificially synthesizing low-resol
ution (LR) images from the high-resolution (HR) ones. However, this dataset prep
aration strategy harms the application of these CNNs in real-world scenarios due
 to the inherent domain gap between the training and testing data. A popular att
empts towards the challenge is unpaired generative adversarial networks, which g
enerate "real" LR counterparts from real HR images using image-to-image translat
ion and then perform super-resolution from "real" LR->SR. Despite great progress
, it is still difficult to synthesize perfect "real" LR images for super-resolut
ion. In this paper, we firstly consider the real-world SR problem from the tradi
tional domain adaptation perspective. We propose a novel unpaired SR training fr
amework based on feature distribution alignment, with which we can obtain degrad
ation-indistinguishable feature maps and then map them to HR images. In order to
 generate better SR images for target LR domain, we introduce several regulariza
tion losses to force the aligned feature to locate around the target domain. Our
 experiments indicate that our SR network obtains the state-of-the-art performan
ce over both blind and unpaired SR methods on diverse datasets.
****

Unaligned Image-to-Image Translation by Learning to Reweight
Shaoan Xie, Mingming Gong, Yanwu Xu, Kun Zhang; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2021, pp. 14174-14184
Unsupervised image-to-image translation aims at learning the mapping from the so
urce to target domain without using paired images for training. An essential yet
 restrictive assumption for unsupervised image translation is that the two domai
ns are aligned, e.g., for the selfie2anime task, the anime (selfie) domain must
contain only anime (selfie) face images that can be translated to some images in
 the other domain. Collecting aligned domains can be laborious and needs lots of
 attention. In this paper, we consider the task of image translation between two
 unaligned domains, which may arise for various possible reasons. To solve this
problem, we propose to select images based on importance reweighting and develop
 a method to learn the weights and perform translation simultaneously and automa
tically. We compare the proposed method with state-of-the-art image translation
approaches and present qualitative and quantitative results on different tasks w
ith unaligned domains. Extensive empirical evidence demonstrates the usefulness
of the proposed problem formulation and the superiority of our method.
****

OSCAR-Net: Object-Centric Scene Graph Attention for Image Attribution
Eric Nguyen, Tu Bui, Viswanathan Swaminathan, John Collomosse; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14499-1
4508
Images tell powerful stories but cannot always be trusted. Matching images back
to trusted sources (attribution) enables users to make a more informed judgment
of the images they encounter online. We propose a robust image hashing algorithm
 to perform such matching. Our hash is sensitive to manipulation of subtle, sali
ent visual details that can substantially change the story told by an image. Yet
 the hash is invariant to benign transformations (changes in quality, codecs, si
zes, shapes, etc.) experienced by images during online redistribution. Our key c
ontribution is OSCAR-Net (Object-centric Scene Graph Attention for Image Attribu
tion Network); a robust image hashing model inspired by recent successes of Tran
sformers in the visual domain. OSCAR-Net constructs a scene graph representation
 that attends to fine-grained changes of every object's visual appearance and th
eir spatial relationships. The network is trained via contrastive learning on a
dataset of original and manipulated images yielding a state of the art image has
h for content fingerprinting that scales to millions of images.
****

A-SDF: Learning Disentangled Signed Distance Functions for Articulated Shape Rep
resentation
Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, Xiaolon
g Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 13001-13011

Recent work has made significant progress on using implicit functions, as a continuous representation for 3D rigid object shape reconstruction. However, much less effort has been devoted to modeling general articulated objects. Compared to rigid objects, articulated objects have higher degrees of freedom, which makes it hard to generalize to unseen shapes. To deal with the large shape variance, we introduce Articulated Signed Distance Functions (A-SDF) to represent articulated shapes with a disentangled latent space, where we have separate codes for encoding shape and articulation. With this disentangled continuous representation, we demonstrate that we can control the articulation input and animate unseen instances with unseen joint angles. Furthermore, we propose a Test-Time Adaptation inference algorithm to adjust our model during inference. We demonstrate our model generalize well to out-of-distribution and unseen data, e.g., partial point clouds and real-world depth images.

****

Consistency-Sensitivity Guided Ensemble Black-Box Adversarial Attacks in Low-Dimensional Spaces

Jianhe Yuan, Zhihai He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7778-7786

Black-box attacks aim to generate adversarial noise tofail the victim deep neural network in the black box. Thecentral task in black-box attack method design is to estimateand characterize the victim model in the high-dimensionalmodel space based on feedback results of queries submittedto the victim network. The central performance goal is tominimize the number of queries needed for successful at-tack. Existing attack methods directly search and refine theadversarial noise in an extremely high-dimensional space,requiring hundreds or even thousands queries to the victimnetwork. To address this challenge, we propose to explore aconsistency and sensitivity guided ensemble attack (CSEA)method in a low-dimensional space. Specifically, we esti-mate the victim model in the black box using a learned lin-ear composition of an ensemble of surrogate models withdiversified network structures. Using random block maskson the input image, these surrogate models jointly constructand submit randomized and sparsified queries to the victimmodel. Based on these query results and guided by a con-sistency constraint, the surrogate models can be trained us-ing a very small number of queries such that their learnedcomposition is able to accurately approximate the victimmodel in the high-dimensional space. The randomized andsparsified queries also provide important information for usto construct an attack sensitivity map for the input image, with which the adversarial attack can be locally refined tofurther increase its success rate. Our extensive experimen-tal results demonstrate that our proposed approach signifi-cantly reduces the number of queries to the victim networkwhile maintaining very high success rates, outperformingexisting black-box attack methods by large margins.

****

Vision Transformer With Progressive Sampling

Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip H.S. Torr, Wayne Zhang, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 387-396

Transformers with powerful global relation modeling abilities have been introduced to fundamental computer vision tasks recently. As a typical example, the Vision Transformer (ViT) directly applies a pure transformer architecture on image classification, by simply splitting images into tokens with a fixed length, and employing transformers to learn relations between these tokens. However, such naive tokenization could destruct object structures, assign grids to uninterested regions such as background, and introduce interference signals. To mitigate the above issues, in this paper, we propose an iterative and progressive sampling strategy to locate discriminative regions. At each iteration, embeddings of the current sampling step are fed into a transformer encoder layer, and a group of sampling offsets is predicted to update the sampling locations for the next step. The progressive sampling is differentiable. When combined with the Vision Transformer, the obtained PS-ViT network can adaptively learn where to look. The proposed PS-ViT is both effective and efficient. When trained from scratch on ImageNet,

PS-ViT performs 3.8% higher than the vanilla ViT in terms of top-1 accuracy with about 4x fewer parameters and 10x fewer FLOPs. Code is available at https://github.com/yuexy/PS-ViT.
****

Exploring Long Tail Visual Relationship Recognition With Large Vocabulary
Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, Mohamed Elhoseiny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15921-15930
Several approaches have been proposed in recent literature to alleviate the long-tail problem, mainly in object classification tasks. In this paper, we make the first large-scale study concerning the task of Long-Tail Visual Relationship Recognition (LTVRR). LTVRR aims at improving the learning of structured visual relationships that come from the long-tail (e.g.,"rabbit grazing on grass"). In this setup, the subject, relation, and object classes each follow a long-tail distribution. To begin our study and make a future benchmark for the community, we introduce two LTVRR-related benchmarks, dubbed VG8K-LT and GQA-LT, built upon the widely used Visual Genome and GQA datasets. We use these benchmarks to study the performance of several state-of-the-art long-tail models on the LTVRR setup. Lastly, we propose a visiolinguistic hubless (VilHub) loss and a Mixup augmentation technique adapted to LTVRR setup, dubbed as RelMix. Both VilHub and RelMix can be easily integrated on top of existing models and despite being simple, our results show that they can remarkably improve the performance, especially on tail classes. Benchmarks, code, and models have been made available at: https://github.com/Vision-CAIR/LTVRR.
****

EM-POSE: 3D Human Pose Estimation From Sparse Electromagnetic Trackers
Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11510-11520
Fully immersive experiences in AR/VR depend on reconstructing the full body pose of the user without restricting their motion. In this paper we study the use of body-worn electromagnetic (EM) field-based sensing for the task of 3D human pose reconstruction. To this end, we present a method to estimate SMPL parameters from 6-12 EM sensors. We leverage a customized wearable system consisting of wireless EM sensors measuring time-synchronized 6D poses at 120 Hz. To provide accurate poses even with little user instrumentation, we adopt a recently proposed hybrid framework, learned gradient descent (LGD), to iteratively estimate SMPL pose and shape from our input measurements. This allows us to harness powerful pose priors to cope with the idiosyncrasies of the input data and achieve accurate pose estimates. The proposed method uses AMASS to synthesize virtual EM-sensor data and we show that it generalizes well to a newly captured real dataset consisting of a total of 36 minutes of motion from 5 subjects. We achieve reconstruction errors as low as 31.8 mm and 13.3 degrees, outperforming both pure learning- and pure optimization-based methods. Code and data is available under https://ait.ethz.ch/projects/2021/em-pose.
****

On Exposing the Challenging Long Tail in Future Prediction of Traffic Actors
Osama Makansi, Özgün Çiçek, Yassine Marrakchi, Thomas Brox; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13147-13157
Predicting the future states of dynamic traffic actors enables autonomous systems to avoid accidents and operate safely. Remarkably, the most critical scenarios are much less frequent and more complex than the uncritical ones. Therefore, uncritical cases dominate the prediction. In this paper, we address specifically the challenging scenarios at the long tail of the dataset distribution. Our analysis shows that the common losses tend to place challenging cases sub-optimally in the embedding space. As a consequence, we propose to supplement the usual loss with a loss that places challenging cases closer to each other in the embedding space. This triggers sharing information among challenging cases and learning specific predictive features. We show on four public datasets that this leads to

improved performance on the hard scenarios while the overall performance stays s
table. The approach is agnostic w.r.t. the used network architecture, input moda
lity or viewpoint, and can be integrated into existing solutions easily.
****

Video Geo-Localization Employing Geo-Temporal Feature Learning and GPS Trajector
y Smoothing
Krishna Regmi, Mubarak Shah; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 12126-12135
In this paper, we address the problem of video geo-localization by proposing a G
eo-Temporal Feature Learning (GTFL) Network to simultaneously learn the discrimi
native features between the query videos and gallery images for estimating the g
eo-spatial trajectory of a query video. Based on a transformer encoder architect
ure, our GTFL model encodes query and gallery data separately, via two dedicated
 branches. The proposed GPS Loss and Clip Triplet Loss exploit the geographical
and temporal proximity between the frames and the clips to jointly learn the que
ry and gallery features. We also propose a deep learning approach to trajectory
smoothing by predicting the outliers in the estimated GPS positions and learning
 the offsets to smooth the trajectory. We build a large dataset from four differ
ent regions of USA; New York, San Francisco, Berkeley and Bay Area using BDD dri
ving videos as query, and by collecting corresponding Google StreetView (GSV) Im
ages for gallery. Extensive evaluations of proposed method on this new dataset a
re provided. Code and dataset details will be made publicly available.
****

ICON: Learning Regular Maps Through Inverse Consistency
Hastings Greer, Roland Kwitt, François-Xavier Vialard, Marc Niethammer; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp
. 3396-3405
Learning maps between data samples is fundamental. Applications range from repre
sentation learning, image translation and generative modeling, to the estimation
 of spatial deformations. Such maps relate feature vectors, or map between featu
re spaces. Well-behaved maps should be regular, which can be imposed explicitly
or may emanate from the data itself. We explore what induces regularity for spat
ial transformations, e.g., when computing image registrations. Classical optimiz
ation-based models compute maps between pairs of samples and rely on an appropri
ate regularizer for well-posedness. Recent deep learning approaches have attempt
ed to avoid using such regularizers altogether by relying on the sample populati
on instead. We explore if it is possible to obtain spatial regularity using an i
nverse consistency loss only and elucidate what explains map regularity in such
a context. We find that deep networks combined with an inverse consistency loss
and randomized off-grid interpolation yield well behaved, approximately diffeomo
rphic, spatial transformations. Despite the simplicity of this approach, our exp
eriments present compelling evidence, on both synthetic and real data, that regu
lar maps can be obtained without carefully tuned explicit regularizers and compe
titive registration performance.
****

ELF-VC: Efficient Learned Flexible-Rate Video Coding
Oren Rippel, Alexander G. Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, L
ubomir Bourdev; Proceedings of the IEEE/CVF International Conference on Computer
 Vision (ICCV), 2021, pp. 14479-14488
While learned video codecs have demonstrated great promise, they have yet to ach
ieve sufficient efficiency for practical deployment. In this work, we propose se
veral ideas for learned video compression which allow for improved performance f
or the low-latency mode (I- and P-frames only) along with a considerable increas
e in computational efficiency. In this setting, for natural videos our approach
compares favorably across the entire R-D curve under metrics PSNR, MS-SSIM and V
MAF against all mainstream video standards (H.264, H.265, AV1) and all ML codecs
. At the same time, our approach runs at least 5x faster and has fewer parameter
s than all ML codecs which report these figures. Our contributions include a fle
xible-rate framework allowing a single model to cover a large and dense range of
 bitrates, at a negligible increase in computation and parameter count; an effic

ient backbone optimized for ML-based codecs; and a novel in-loop flow prediction scheme which leverages prior information towards more efficient compression. We benchmark our method, which we call ELF-VC (Efficient, Learned and Flexible Video Coding) on popular video test sets UVG and MCL-JCV under metrics PSNR, MS-SSIM and VMAF. For example, on UVG under PSNR, it reduces the BD-rate by 44% against H.264, 26% against H.265, 15% against AV1, 35% against the current best ML codec. At the same time, on an NVIDIA Titan V GPU our approach encodes/decodes VGA at 49/91 FPS, HD 720 at 19/35 FPS, and HD 1080 at 10/18 FPS.
****

Structure-Preserving Deraining With Residue Channel Prior Guidance
Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, Tieyong Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4238-4247
Single image deraining is important for many high-level computer vision tasks since the rain streaks can severely degrade the visibility of images, thereby affecting the recognition and analysis of the image. Recently, many CNN-based methods have been proposed for rain removal. Although these methods can remove part of the rain streaks, it is difficult for them to adapt to real-world scenarios and restore high-quality rain-free images with clear and accurate structures. To solve this problem, we propose a Structure-Preserving Deraining Network (SPDNet) with RCP guidance. SPDNet directly generates high-quality rain-free images with clear and accurate structures under the guidance of RCP but does not rely on any rain-generating assumptions. Specifically, we found that the RCP of images contains more accurate structural information than rainy images. Therefore, we introduced it to our deraining network to protect structure information of the rain-free image. Meanwhile, a Wavelet-based Multi-Level Module (WMLM) is proposed as the backbone for learning the background information of rainy images and an Interactive Fusion Module (IFM) is designed to make full use of RCP information. In addition, an iterative guidance strategy is proposed to gradually improve the accuracy of RCP, refining the result in a progressive path. Extensive experimental results on both synthetic and real-world datasets demonstrate that the proposed model achieves new state-of-the-art results. Code: https://github.com/Joyies/SPDNet
****

Fast and Efficient DNN Deployment via Deep Gaussian Transfer Learning
Qi Sun, Chen Bai, Tinghuan Chen, Hao Geng, Xinyun Zhang, Yang Bai, Bei Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5380-5390
Deep neural networks (DNNs) have been widely used recently while their hardware deployment optimizations are very time-consuming and the historical deployment knowledge is not utilized efficiently. In this paper, to accelerate the optimization process and find better deployment configurations, we propose a novel transfer learning method based on deep Gaussian processes (DGPs). Firstly, a deep Gaussian process (DGP) model is built on the historical data to learn empirical knowledge. Secondly, to transfer knowledge to a new task, a tuning set is sampled for the new task under the guidance of the DGP model. Then DGP is tuned according to the tuning set via maximum-a-posteriori (MAP) estimation to accommodate for the new task and finally used to guide the deployments of the task. The experiments show that our method achieves the best inference latencies of convolutions while accelerating the optimization process significantly, compared with previous arts.
****

Towards Complete Scene and Regular Shape for Distortion Rectification by Curve-Aware Extrapolation
Kang Liao, Chunyu Lin, Yunchao Wei, Feng Li, Shangrong Yang, Yao Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14569-14578
The wide-angle lens gains increasing attention since it can capture a wide field-of-view scene (FoV). However, the obtained image is contaminated with radial distortion, making the scene not realistic. Previous distortion rectification meth

ods rectify the image in a rectangle or invagination, failing to display the com
plete content and regular shape simultaneously. In this paper, we rethink the re
presentation of rectification results and present a Rectification OutPainting (R
OP) method, aiming to extrapolate the coherent semantics to the blank area and c
reate a wider FoV beyond the original wide-angle lens. To address the specific c
hallenges such as the variable painting region and curve boundary, a rectificati
on module is designed to rectify the image with geometry supervision, and the ex
trapolated results are generated using a dual conditional expansion strategy. In
 terms of the spatially discounted correlation, a curve-aware correlation measur
ement is proposed to focus on the generated region to enforce the local consiste
ncy. To our knowledge, we are the first to tackle the challenging rectification
 via outpainting, and our curve-aware strategy can reach a rectification construc
tion with complete content and regular shape. Extensive experiments well demonst
rate the superiority of our ROP over other state-of-the-art solutions.
****

ViewNet: Unsupervised Viewpoint Estimation From Conditional Generation
Octave Mariotti, Oisin Mac Aodha, Hakan Bilen; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 10418-10428
Understanding the 3D world without supervision is currently a major challenge in
 computer vision as the annotations required to supervise deep networks for task
s in this domain are expensive to obtain on a large scale. In this paper, we add
ress the problem of unsupervised viewpoint estimation. We formulate this as a se
lf-supervised learning task, where image reconstruction provides the supervision
 needed to predict the camera viewpoint. Specifically, we make use of pairs of i
mages of the same object at training time, from unknown viewpoints, to self-supe
rvise training by combining the viewpoint information from one image with the ap
pearance information from the other. We demonstrate that using a perspective spa
tial transformer allows efficient viewpoint learning, outperforming existing uns
upervised approaches on synthetic data, and obtains competitive results on the c
hallenging PASCAL3D+ dataset.
****

High-Resolution Optical Flow From 1D Attention and Correlation
Haofei Xu, Jiaolong Yang, Jianfei Cai, Juyong Zhang, Xin Tong; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10498-1
0507
Optical flow is inherently a 2D search problem, and thus the computational compl
exity grows quadratically with respect to the search window, making large displa
cements matching infeasible for high-resolution images. In this paper, we take i
nspiration from Transformers and propose a new method for high-resolution optica
l flow estimation with significantly less computation. Specifically, a 1D attent
ion operation is first applied in the vertical direction of the target image, an
d then a simple 1D correlation in the horizontal direction of the attended image
 is able to achieve 2D correspondence modeling effect. The directions of attenti
on and correlation can also be exchanged, resulting in two 3D cost volumes that
are concatenated for optical flow estimation. The novel 1D formulation empowers
our method to scale to very high-resolution input images while maintaining compe
titive performance. Extensive experiments on Sintel, KITTI and real-world 4K (21
60 x 3840) resolution images demonstrated the effectiveness and superiority of o
ur proposed method. Code and models are available at https://github.com/haofeixu
/flow1d.
****

RGB-D Saliency Detection via Cascaded Mutual Information Minimization
Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, Ling Sh
ao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2021, pp. 4338-4347
Existing RGB-D saliency detection models do not explicitly encourage RGB and dep
th to achieve effective multi-modal learning. In this paper, we introduce a nove
l multi-stage cascaded learning framework via mutual information minimization to
 explicitly model the multi-modal information between RGB image and depth data.
Specifically, we first map the feature of each mode to a lower dimensional featu

re vector, and adopt mutual information minimization as a regularizer to reduce the redundancy between appearance features from RGB and geometric features from depth. We then perform multi-stage cascaded learning to impose the mutual information minimization constraint at every stage of the network. Extensive experiments on benchmark RGB-D saliency datasets illustrate the effectiveness of our framework. Further, to prosper the development of this field, we contribute the largest (7x larger than NJU2K) COME20K dataset, which contains 15,625 image pairs with high quality polygon-/scribble-/object-/instance-/rank-level annotations. Based on these rich labels, we additionally construct four new benchmarks (Code, results, and benchmarks will be made publicly available.) with strong baselines and observe some interesting phenomena, which can motivate future model design.
****

A Weakly Supervised Amodal Segmenter With Boundary Uncertainty Estimation
Khoi Nguyen, Sinisa Todorovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7396-7405
This paper addresses weakly supervised amodal instance segmentation, where the goal is to segment both visible and occluded (amodal) object parts, while training provides only ground-truth visible (modal) segmentations. Following prior work, we use data manipulation to generate occlusions in training images and thus train a segmenter to predict amodal segmentations of the manipulated data. The resulting predictions on training images are taken as the pseudo-ground truth for the standard training of Mask-RCNN, which we use for amodal instance segmentation of test images. For generating the pseudo-ground truth, we specify a new Amodal Segmenter based on Boundary Uncertainty estimation (ASBU) and make two contributions. First, while prior work uses the occluder's mask, our ASBU uses the occlusion boundary as input. Second, ASBU estimates an uncertainty map of the prediction. The estimated uncertainty regularizes learning such that lower segmentation loss is incurred on regions with high uncertainty. ASBU achieves significant performance improvement relative to the state of the art on the COCOA and KINS datasets in three tasks: amodal instance segmentation, amodal completion, and ordering recovery.
****

Cross-Camera Convolutional Color Constancy
Mahmoud Afifi, Jonathan T. Barron, Chloe LeGendre, Yun-Ta Tsai, Francois Bleibel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1981-1990
We present "Cross-Camera Convolutional Color Constancy" (C5), a learning-based method, trained on images from multiple cameras, that accurately estimates a scene's illuminant color from raw images captured by a new camera previously unseen during training. C5 is a hypernetwork-like extension of the convolutional color constancy (CCC) approach: C5 learns to generate the weights of a CCC model that is then evaluated on the input image, with the CCC weights dynamically adapted to different input content. Unlike prior cross-camera color constancy models, which are usually designed to be agnostic to the spectral properties of test-set images from unobserved cameras, C5 approaches this problem through the lens of transductive inference: additional unlabeled images are provided as input to the model at test time, which allows the model to calibrate itself to the spectral properties of the test-set camera during inference. C5 achieves state-of-the-art accuracy for cross-camera color constancy on several datasets, is fast to evaluate (7 and 90 ms per image on a GPU or CPU, respectively), and requires little memory (2 MB), and thus is a practical solution to the problem of calibration-free automatic white balance for mobile photography.
****

Kernel Methods in Hyperbolic Spaces
Pengfei Fang, Mehrtash Harandi, Lars Petersson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10665-10674
Embedding data in hyperbolic spaces has proven beneficial for many advanced machine learning applications such as image classification and word embeddings. However, working in hyperbolic spaces is not without difficulties as a result of its curved geometry (e.g., computing the Frechet mean of a set of points requires a

n iterative algorithm). Furthermore, in Euclidean spaces, one can resort to kern el machines that not only enjoy rich theoretical properties but that can also le ad to superior representational power (e.g., infinite-width neural networks). In this paper, we introduce positive definite kernel functions for hyperbolic spac es. This brings in two major advantages, 1. kernelization will pave the way to s eamlessly benefit from kernel machines in conjunction with hyperbolic embeddings , and 2. the rich structure of the Hilbert spaces associated with kernel machine s enables us to simplify various operations involving hyperbolic data. That said , identifying valid kernel functions on curved spaces is not straightforward and is indeed considered an open problem in the learning community. Our work addres ses this gap and develops several valid positive definite kernels in hyperbolic spaces, including the universal ones (e.g., RBF). We comprehensively study the p roposed kernels on a variety of challenging tasks including few-shot learning, z ero-shot learning, person re-identification and knowledge distillation, showing the superiority of the kernelization for hyperbolic representations.

****

Towards Alleviating the Modeling Ambiguity of Unsupervised Monocular 3D Human Po se Estimation

Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, Wenjun Zhang; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 021, pp. 8651-8660

In this work, we study the ambiguity problem in the task of unsupervised 3D huma n pose estimation from 2D counterpart. On one hand, without explicit annotation, the scale of 3D pose is difficult to be accurately captured (scale ambiguity). On the other hand, one 2D pose might correspond to multiple 3D gestures, where t he lifting procedure is inherently ambiguous (pose ambiguity). Previous methods generally use temporal constraints (e.g., constant bone length and motion smooth ness) to alleviate the above issues. However, these methods commonly enforce the outputs to fulfill multiple training objectives simultaneously, which often lea d to sub-optimal results. In contrast to the majority of previous works, we prop ose to split the whole problem into two sub-tasks, i.e., optimizing 2D input pos es via a scale estimation module and then mapping optimized 2D pose to 3D counte rpart via a pose lifting module. Furthermore, two temporal constraints are propo sed to alleviate the scale and pose ambiguity respectively. These two modules ar e optimized via a iterative training scheme with corresponding temporal constrai nts, which effectively reduce the learning difficulty and lead to better perform ance. Results on the Human3.6M dataset demonstrate that our approach improves up on the prior art by 23.1% and also outperforms several weakly supervised approac hes that rely on 3D annotations. Our project is available at https://sites.googl e.com/view/ambiguity-aware-hpe.

****

Geometric Deep Neural Network Using Rigid and Non-Rigid Transformations for Huma n Action Recognition

Rasha Friji, Hassen Drira, Faten Chaieb, Hamza Kchok, Sebastian Kurtek; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 12611-12620

Deep Learning architectures, albeit successful in mostcomputer vision tasks, wer e designed for data with an un-derlying Euclidean structure, which is not usuall y fulfilledsince pre-processed data may lie on a non-linear space.In this paper, we propose a geometry aware deep learn-ing approach using rigid and non rigid t ransformation opti-mization for skeleton-based action recognition. Skeleton se-q uences are first modeled as trajectories on Kendall's shapespace and then mapped to the linear tangent space. The re-sulting structured data are then fed to a d eep learning archi-tecture, which includes a layer that optimizes over rigid and non rigid transformations of the 3D skeletons, followed bya CNN-LSTM network. Th e assessment on two large scaleskeleton datasets, namely NTU-RGB+D and NTU-RGB+D 120, has proven that the proposed approach outperformsexisting geometric deep le arning methods and exceeds re-cently published approaches with respect to the ma jority of configurations.

****

Enhanced Boundary Learning for Glass-Like Object Segmentation
Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, LuBin Weng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15859-15868
Glass-like objects such as windows, bottles, and mirrors exist widely in the real world. Sensing these objects has many applications, including robot navigation and grasping. However, this task is very challenging due to the arbitrary scenes behind glass-like objects. This paper aims to solve the glass-like object segmentation problem via enhanced boundary learning. In particular, we first propose a novel refined differential module that outputs finer boundary cues. We then introduce an edge-aware point-based graph convolution network module to model the global shape along the boundary. We use these two modules to design a decoder that generates accurate and clean segmentation results, especially on the object contours. Both modules are lightweight and effective: they can be embedded into various segmentation models. In extensive experiments on three recent glass-like object segmentation datasets, including Trans10k, MSD, and GDD, our approach establishes new state-of-the-art results. We also illustrate the strong generalization properties of our method on three generic segmentation datasets, including Cityscapes, BDD, and COCO Stuff. Code and models will be available for further research.
****
Self-Supervised Pretraining of 3D Features on Any Point-Cloud
Zaiwei Zhang, Rohit Girdhar, Armand Joulin, Ishan Misra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10252-10263
Pretraining on large labeled datasets is a prerequisite to achieve good performance in many computer vision tasks like image recognition, video understanding etc. However, pretraining is not widely used for 3D recognition tasks where state-of-the-art methods train models from scratch. A primary reason is the lack of large annotated datasets because 3D data labelling is time-consuming. Recent work shows that self-supervised learning is useful to pretrain models in 3D but requires multi-view data and point correspondences. We present a simple self-supervised pretraining method that can work with single-view depth scans acquired by varied sensors, without 3D registration and point correspondences. We pretrain standard point cloud and voxel based model architectures, and show that joint pretraining further improves performance. We evaluate our models on 9 benchmarks for object detection, semantic segmentation, and object classification, where they achieve state-of-the-art results. Most notably, we set a new state-of-the-art for object detection on ScanNet (69.0% mAP) and SUNRGBD (63.5% mAP). Our pretrained models are label efficient and improve performance for classes with few examples.
****
N-ImageNet: Towards Robust, Fine-Grained Object Recognition With Event Cameras
Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, Young Min Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2146-2156
We introduce N-ImageNet, a large-scale dataset targeted for robust, fine-grained object recognition with event cameras. The dataset is collected using programmable hardware in which an event camera consistently moves around a monitor displaying images from ImageNet. N-ImageNet serves as a challenging benchmark for event-based object recognition, due to its large number of classes and samples. We empirically show that pretraining on N-ImageNet improves the performance of event-based classifiers and helps them learn with few labeled data. In addition, we present several variants of N-ImageNet to test the robustness of event-based classifiers under diverse camera trajectories and severe lighting conditions, and propose a novel event representation to alleviate the performance degradation. To the best of our knowledge, we are the first to quantitatively investigate the consequences caused by various environmental conditions on event-based object recognition algorithms. N-ImageNet and its variants are expected to guide practical implementations for deploying event-based object recognition algorithms in the real world.

****
Diagonal Attention and Style-Based GAN for Content-Style Disentanglement in Image Generation and Translation
Gihyun Kwon, Jong Chul Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13980-13989
One of the important research topics in image generative models is to disentangle the spatial contents and styles for their separate control. Although StyleGAN can generate content feature vectors from random noises, the resulting spatial content control is primarily intended for minor spatial variations, and the disentanglement of global content and styles is by no means complete. Inspired by a mathematical understanding of normalization and attention, here we present a novel hierarchical adaptive Diagonal spatial ATtention (DAT) layers to separately manipulate the spatial contents from styles in a hierarchical manner. Using DAT and AdaIN, our method enables coarse-to-fine level disentanglement of spatial contents and styles. In addition, our generator can be easily integrated into the GAN inversion framework so that the content and style of translated images from multi-domain image translation tasks can be flexibly controlled. By using various datasets, we confirm that the proposed method not only outperforms the existing models in disentanglement scores, but also provides more flexible control over spatial features in the generated images.
****
Who's Waldo? Linking People Across Text and Images
Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, Hadar Averbuch-Elor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1374-1384
We present a task and benchmark dataset for person-centric visual grounding, the problem of linking between people named in a caption and people pictured in an image. In contrast to prior work in visual grounding, which is predominantly object-based, our new task masks out the names of people in captions in order to encourage methods trained on such image--caption pairs to focus on contextual cues (such as rich interactions between multiple people), rather than learning associations between names and appearances. To facilitate this task, we introduce a new dataset, Who's Waldo, mined automatically from image--caption data on Wikimedia Commons. We propose a Transformer-based method that outperforms several strong baselines on this task, and are releasing our data to the research community to spur work on contextual models that consider both vision and language.
****
Switchable K-Class Hyperplanes for Noise-Robust Representation Learning
Boxiao Liu, Guanglu Song, Manyuan Zhang, Haihang You, Yu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3019-3028
Optimizing the K-class hyperplanes in the latent space has become the standard paradigm for efficient representation learning. However, it's almost impossible to find an optimal K-class hyperplane to accurately describe the latent space of massive noisy data. For this potential problem, we constructively propose a new method, named Switchable K-class Hyperplanes (SKH), to sufficiently describe the latent space by the mixture of K-class hyperplanes. It can directly replace the conventional single K-class hyperplane optimization as the new paradigm for noise-robust representation learning. When collaborated with the popular ArcFace on million-level data representation learning, we found that the switchable manner in SKH can effectively eliminate the gradient conflict generated by real-world label noise on a single K-class hyperplane. Moreover, combined with the margin-based loss functions (e.g. ArcFace), we propose a simple Posterior Data Clean strategy to reduce the model optimization deviation on clean dataset caused by the reduction of valid categories in each K-class hyperplane. Extensive experiments demonstrate that the proposed SKH easily achieves new state-of-the-art on IJB-B and IJB-C by encouraging noise-robust representation learning.
****
Transformer-Based Attention Networks for Continuous Pixel-Wise Prediction
Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, Elisa Ricci; Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16269-16279
While convolutional neural networks have shown a tremendous impact on various computer vision tasks, they generally demonstrate limitations in explicitly modeling long-range dependencies due to the intrinsic locality of the convolution operation. Initially designed for natural language processing tasks, Transformers have emerged as alternative architectures with innate global self-attention mechanisms to capture long-range dependencies. In this paper, we propose TransDepth, an architecture that benefits from both convolutional neural networks and transformers. To avoid the network losing its ability to capture local-level details due to the adoption of transformers, we propose a novel decoder that employs attention mechanisms based on gates. Notably, this is the first paper that applies transformers to pixel-wise prediction problems involving continuous labels (i.e., monocular depth prediction and surface normal estimation). Extensive experiments demonstrate that the proposed TransDepth achieves state-of-the-art performance on three challenging datasets. Our code is available at: https://github.com/ygjwd12345/TransDepth.
****

BlockPlanner: City Block Generation With Vectorized Graph Representation
Linning Xu, Yuanbo Xiangli, Anyi Rao, Nanxuan Zhao, Bo Dai, Ziwei Liu, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5077-5086
City modeling is the foundation for computational urban planning, navigation, and entertainment. In this work, we present the first generative model of city blocks named BlockPlanner, and showcase its ability to synthesize valid city blocks with varying land lots configurations. We propose a novel vectorized city block representation utilizing a ring topology and a two-tier graph to capture the global and local structures of a city block. Each land lot is abstracted into a vector representation covering both its 3D geometry and land use semantics. Such vectorized representation enables us to deploy a lightweight network to capture the underlying distribution of land lots configuration in a city block. To enforce intrinsic spatial constraints of a valid city block, a set of effective loss functions are imposed to shape rational results. We contribute a pilot city block dataset to demonstrate the effectiveness and efficiency of our representation and framework over the state-of-the-art. Notably, our BlockPlanner is also able to edit and manipulate city blocks, enabling several useful applications, e.g., topology refinement and footprint generation.
****

PCAM: Product of Cross-Attention Matrices for Rigid Registration of Point Clouds
Anh-Quan Cao, Gilles Puy, Alexandre Boulch, Renaud Marlet; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13229-13238
Rigid registration of point clouds with partial overlaps is a longstanding problem usually solved in two steps: (a) finding correspondences between the point clouds; (b) filtering these correspondences to keep only the most reliable ones to estimate the transformation. Recently, several deep nets have been proposed to solve these steps jointly. We built upon these works and propose PCAM: a neural network whose key element is a pointwise product of cross-attention matrices that permits to mix both low-level geometric and high-level contextual information to find point correspondences. These cross-attention matrices also permits the exchange of context information between the point clouds, at each layer, allowing the network construct better matching features within the overlapping regions. The experiments show that PCAM achieves state-of-the-art results among methods which, like us, solve steps (a) and (b) jointly via deepnets.
****

CCT-Net: Category-Invariant Cross-Domain Transfer for Medical Single-to-Multiple Disease Diagnosis
Yi Zhou, Lei Huang, Tao Zhou, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8260-8270
A medical imaging model is usually explored for the diagnosis of a single disease. However, with the expanding demand for multi-disease diagnosis in clinical ap

plications, multi-function solutions need to be investigated. Previous works pro posed to either exploit different disease labels to conduct transfer learning th rough fine-tuning, or transfer knowledge across different domains with similar d iseases. However, these methods still cannot address the real clinical challenge - a multi-disease model is required but annotations for each disease are not al ways available. In this paper, we introduce the task of transferring knowledge f rom single-disease diagnosis (source domain) to enhance multi-disease diagnosis (target domain). A category-invariant cross-domain transfer (CCT) method is prop osed to address this single-to-multiple extension. First, for domain-specific ta sk learning, we present a confidence weighted pooling (CWP) to obtain coarse hea tmaps for different disease categories. Then, conditioned on these heatmaps, cat egory-invariant feature refinement (CIFR) blocks are proposed to better localize discriminative semantic regions related to the corresponding diseases. The cate gory-invariant characteristic enables transferability from the source domain to the target domain. We validate our method in two popular areas: extending diabet ic retinopathy to identifying multiple ocular diseases, and extending glioma ide ntification to the diagnosis of other brain tumors.

****

FLAR: A Unified Prototype Framework for Few-Sample Lifelong Active Recognition
Lei Fan, Peixi Xiong, Wei Wei, Ying Wu; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2021, pp. 15394-15403

Intelligent agents with visual sensors are allowed to actively explore their obs ervations for better recognition performance. This task is referred to as Active Recognition (AR). Currently, most methods toward AR are implemented under a fix ed-category setting, which constrains their applicability in realistic scenarios that need to incrementally learn new classes without retraining from scratch. F urther, collecting massive data for novel categories is expensive. To address th is demand, in this paper, we propose a unified framework towards Few-sample Life long Active Recognition (FLAR), which aims at performing active recognition on p rogressively arising novel categories that only have few training samples. Three difficulties emerge with FLAR: the lifelong recognition policy learning, the kn owledge preservation of old categories, and the lack of training samples. To thi s end, our approach integrates prototypes, a robust representation for limited t raining samples, into a reinforcement learning solution, which motivates the age nt to move towards views resulting in more discriminative features. Catastrophic forgetting during lifelong learning is then alleviated with knowledge distillat ion. Extensive experiments across two datasets, respectively for object and scen e recognition, demonstrate that even without large training samples, the propose d approach could learn to actively recognize novel categories in a class-increme ntal behavior.

****

VLGrammar: Grounded Grammar Induction of Vision and Language
Yining Hong, Qing Li, Song-Chun Zhu, Siyuan Huang; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2021, pp. 1665-1674

Cognitive grammar suggests that the acquisition of language grammar is grounded within visual structures. While grammar is an essential representation of natura l language, it also exists ubiquitously in vision to represent the hierarchical part-whole structure. In this work, we study grounded grammar induction of visio n and language in a joint learning framework. Specifically, we present VLGrammar , a method that uses compound probabilistic context-free grammars (compound PCFG s) to induce the language grammar and the image grammar simultaneously. We propo se a novel contrastive learning framework to guide the joint learning of both mo dules. To provide a benchmark for the grounded grammar induction task, we collec t a large-scale dataset, PartIt, which contains human-written sentences that des cribe part-level semantics for 3D objects. Experiments on the PartIt dataset sho w that VLGrammar outperforms all baselines in image grammar induction and langua ge grammar induction. The learned VLGrammar naturally benefits related downstrea m tasks. Specifically, it improves the image unsupervised clustering accuracy by 30%, and performs well in image retrieval and text retrieval. Notably, the indu ced grammar shows superior generalizability by easily generalizing to unseen cat

egories.
****
Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning
Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, Xiaolong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9062-9071
Meta-learning has been the most common framework for few-shot learning in recent years. It learns the model from collections of few-shot classification tasks, which is believed to have a key advantage of making the training objective consistent with the testing objective. However, some recent works report that by training for whole-classification, i.e. classification on the whole label-set, it can get comparable or even better embedding than many meta-learning algorithms. The edge between these two lines of works has yet been underexplored, and the effectiveness of meta-learning in few-shot learning remains unclear. In this paper, we explore a simple process: meta-learning over a whole-classification pre-trained model on its evaluation metric. We observe this simple method achieves competitive performance to state-of-the-art methods on standard benchmarks. Our further analysis shed some light on understanding the trade-offs between the meta-learning objective and the whole-classification objective in few-shot learning.
****
CPFN: Cascaded Primitive Fitting Networks for High-Resolution Point Clouds
Eric-Tuan Lê, Minhyuk Sung, Duygu Ceylan, Radomir Mech, Tamy Boubekeur, Niloy J. Mitra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7457-7466
Representing human-made objects as a collection of base primitives has a long history in computer vision and reverse engineering. In the case of high-resolution point cloud scans, the challenge is to be able to detect both large primitives as well as those explaining the detailed parts. While the classical RANSAC approach requires case-specific parameter tuning, state-of-the-art networks are limited by memory consumption of their backbone modules such as PointNet++, and hence fail to detect the fine-scale primitives. We present Cascaded Primitive Fitting Networks (CPFN) that relies on an adaptive patch sampling network to assemble detection results of global and local primitive detection networks. As a key enabler, we present a merging formulation that dynamically aggregates the primitives across global and local scales. Our evaluation demonstrates that CPFN improves the state-of-the-art SPFN performance by 13-14% on high-resolution point cloud datasets and specifically improves the detection of fine-scale primitives by 20-22%. Our code is available at: https://github.com/erictuanle/CPFN
****
PARTS: Unsupervised Segmentation With Slots, Attention and Independence Maximization
Daniel Zoran, Rishabh Kabra, Alexander Lerchner, Danilo J. Rezende; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10439-10447
From an early age, humans perceive the visual world as composed of coherent objects with distinctive properties such as shape, size, and color. There is great interest in building models that are able to learn similar structure, ideally in an unsupervised manner. Learning such structure from complex 3D scenes that include clutter, occlusions, interactions, and camera motion is still an open challenge. We present a model that is able to segment visual scenes from complex 3D environments into distinct objects, learn disentangled representations of individual objects, and form consistent and coherent predictions of future frames, in a fully unsupervised manner. Our model (named PARTS) builds on recent approaches that utilize iterative amortized inference and transition dynamics for deep generative models. We achieve dramatic improvements in performance by introducing several novel contributions. We introduce a recurrent slot-attention like encoder which allows for top-down influence during inference. Unlike prior work, we eschew using an auto-regressive prior when modeling image sequences, and demonstrate that a fixed frame-independent prior is superior for the purpose of scene segmentation and representation learning. We demonstrate our model's success on three

different video datasets (the popular benchmark CLEVRER; a simulated 3D Playroom environment; and a real-world Robotics Arm dataset). Finally, we analyze the contributions of the various model components and the representations learned by the model.
****

Fine-Grained Semantics-Aware Representation Enhancement for Self-Supervised Monocular Depth Estimation
Hyunyoung Jung, Eunhyeok Park, Sungjoo Yoo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12642-12652
Self-supervised monocular depth estimation has been widely studied, owing to its practical importance and recent promising improvements. However, most works suffer from limited supervision of photometric consistency, especially in weak texture regions and at object boundaries. To overcome this weakness, we propose novel ideas to improve self-supervised monocular depth estimation by leveraging cross-domain information, especially scene semantics. We focus on incorporating implicit semantic knowledge into geometric representation enhancement and suggest two ideas: a metric learning approach that exploits the semantics-guided local geometry to optimize intermediate depth representations and a novel feature fusion module that judiciously utilizes cross-modality between two heterogeneous feature representations. We comprehensively evaluate our methods on the KITTI dataset and demonstrate that our method outperforms state-of-the-art methods. The source code is available at https://github.com/hyBlue/FSRE-Depth.
****

Learning Signed Distance Field for Multi-View Surface Reconstruction
Jingyang Zhang, Yao Yao, Long Quan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6525-6534
Recent works on implicit neural representations have shown promising results for multi-view surface reconstruction. However, most approaches are limited to relatively simple geometries and usually require clean object masks for reconstructing complex and concave objects. In this work, we introduce a novel neural surface reconstruction framework that leverages the knowledge of stereo matching and feature consistency to optimize the implicit surface representation. More specifically, we apply a signed distance field (SDF) and a surface light field to represent the scene geometry and appearance respectively. The SDF is directly supervised by geometry from stereo matching, and is refined by optimizing the multi-view feature consistency and the fidelity of rendered images. Our method is able to improve the robustness of geometry estimation and support reconstruction of complex scene topologies. Extensive experiments have been conducted on DTU, EPFL and Tanks and Temples datasets. Compared to previous state-of-the-art methods, our method achieves better mesh reconstruction in wide open scenes without masks as input.
****

Pose Correction for Highly Accurate Visual Localization in Large-Scale Indoor Spaces
Janghun Hyeon, Joohyung Kim, Nakju Doh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15974-15983
Indoor visual localization is significant for various applications such as autonomous robots, augmented reality, and mixed reality. Recent advances in visual localization have demonstrated their feasibility in large-scale indoor spaces through coarse-to-fine methods that typically employ three steps: image retrieval, pose estimation, and pose selection. However, further research is needed to improve the accuracy of large-scale indoor visual localization. We demonstrate that the limitations in the previous methods can be attributed to the sparsity of image positions in the database, which causes view-differences between a query and a retrieved image from the database. In this paper, to address this problem, we propose a novel module, named pose correction, that enables re-estimation of the pose with local feature matching in a similar view by reorganizing the local features. This module enhances the accuracy of the initially estimated pose and assigns more reliable ranks. Furthermore, the proposed method achieves a new state-of-the-art performance with an accuracy of more than 90% within 1.0m in the chal

lenging indoor benchmark dataset InLoc for the first time.
****

A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction
Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, Guiqing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13588-13597
In this paper, we propose HF2-VAD, a Hybrid framework that integrates Flow reconstruction and Frame prediction seamlessly to handle Video Anomaly Detection. Firstly, we design the network of ML-MemAE-SC (Multi-Level Memory modules in an Autoencoder with Skip Connections) to memorize normal patterns for optical flow reconstruction so that abnormal events can be sensitively identified with larger flow reconstruction errors. More importantly, conditioned on the reconstructed flows, we then employ a Conditional Variational Autoencoder (CVAE), which captures the high correlation between video frame and optical flow, to predict the next frame given several previous frames. By CVAE, the quality of flow reconstruction essentially influences that of frame prediction. Therefore, poorly reconstructed optical flows of abnormal events further deteriorate the quality of the final predicted future frame, making the anomalies more detectable. Experimental results demonstrate the effectiveness of the proposed method. Code is available at https://github.com/LiUzHiAn/hf2vad.
****

PointBA: Towards Backdoor Attacks in 3D Point Cloud
Xinke Li, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, Joey Tianyi Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16492-16501
3D deep learning has been increasingly more popular for a variety of tasks including many safety-critical applications. However, recently several works raise the security issues of 3D deep models. Although most of them consider adversarial attacks, we identify that backdoor attack is indeed a more serious threat to 3D deep learning systems but remains unexplored. We present the backdoor attacks in 3D point cloud with a unified framework that exploits the unique properties of 3D data and networks. In particular, we design two attack approaches on point cloud: the poison-label backdoor attack (PointPBA) and the clean-label backdoor attack (PointCBA). The first one is straightforward and effective in practice, while the latter is more sophisticated assuming there are certain data inspections. The attack algorithms are mainly motivated and developed by 1) the recent discovery of 3D adversarial samples suggesting the vulnerability of deep models under spatial transformation; 2) the proposed feature disentanglement technique that manipulates the feature of the data through optimization methods and its potential to embed a new task. Extensive experiments show the efficacy of the PointPBA with over 95% success rate across various 3D datasets and models, and the more stealthy PointCBA with around 50% success rate. Our proposed backdoor attack in 3D point cloud is expected to perform as a baseline for improving the robustness of 3D deep models.
****

Linguistically Routing Capsule Network for Out-of-Distribution Visual Question Answering
Qingxing Cao, Wentao Wan, Keze Wang, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1614-1623
Generalization on out-of-distribution (OOD) test data is an essential but underexplored topic in visual question answering. Current state-of-the-art VQA models often exploit the biased correlation between data and labels, which results in a large performance drop when the test and training data have different distributions. Inspired by the fact that humans can recognize novel concepts by composing existed concepts and capsule network's ability of representing part-whole hierarchies, we propose to use capsules to represent parts and introduce "Linguistically Routing" to merge parts with human-prior hierarchies. Specifically, we first fuse visual features with a single question word as atomic parts. Then we intro

duce the "Linguistically Routing" to reweight the capsule connections between tw
o layers such that: 1) the lower layer capsules can transfer their outputs to th
e most compatible higher capsules, and 2) two capsules can be merged if their co
rresponding words are merged in the question parse tree. The routing process max
imizes the above unary and binary potentials across multiple layers and finally
carves a tree structure inside the capsule network. We evaluate our proposed rou
ting method on the CLEVR compositional generation test, the VQA-CP2 dataset and
the VQAv2 dataset. The experimental results show that our proposed method can im
prove current VQA models on OOD split without losing performance on the in-domai
n test data.
****

Neural Articulated Radiance Field
Atsuhiro Noguchi, Xiao Sun, Stephen Lin, Tatsuya Harada; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5762-5772
We present Neural Articulated Radiance Field (NARF), a novel deformable 3D repre
sentation for articulated objects learned from images. While recent advances in
3D implicit representation have made it possible to learn models of complex obje
cts, learning pose-controllable representations of articulated objects remains a
challenge, as current methods require 3D shape supervision and are unable to re
nder appearance. In formulating an implicit representation of 3D articulated obj
ects, our method considers only the rigid transformation of the most relevant ob
ject part in solving for the radiance field at each 3D location. In this way, th
e proposed method represents pose-dependent changes without significantly increa
sing the computational complexity. NARF is fully differentiable and can be train
ed from images with pose annotations. Moreover, through the use of an autoencode
r, it can learn appearance variations over multiple instances of an object class
. Experiments show that the proposed method is efficient and can generalize well
to novel poses. The code is available for research purposes at https://github.c
om/nogu-atsu/NARF
****

Region Similarity Representation Learning
Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, Trevor Darrell; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 10539-10548
We present Region Similarity Representation Learning (ReSim), a new approach to
self-supervised representation learning for localization-based tasks such as obj
ect detection and segmentation. While existing work has largely focused on learn
ing global representations for an entire image, ReSim learns both regional repre
sentations for localization as well as semantic image-level representations. ReS
im operates by sliding a fixed-sized window across the overlapping area between
two views (e.g., image crops), aligning these areas with their corresponding con
volutional feature map regions, and then maximizing the feature similarity acros
s views. As a result, ReSim learns spatially and semantically consistent feature
representation throughout the convolutional feature maps of a neural network. A
shift or scale of an image region, e.g., a shift or scale of an object, has a c
orresponding change in the feature maps; this allows downstream tasks to leverag
e these representations for localization. Through object detection, instance seg
mentation, and dense pose estimation experiments, we illustrate how ReSim learns
representations which significantly improve the localization and classification
performance compared to a competitive MoCo-v2 baseline: +2:7 APbb75 VOC, +1:1 A
P75 COCO, and +1:9 APmk Cityscapes. We will release our code and pre-trained mod
els.
****

Learning of Visual Relations: The Devil Is in the Tails
Alakh Desai, Tz-Ying Wu, Subarna Tripathi, Nuno Vasconcelos; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15404-154
13
Significant effort has been recently devoted to modeling visual relations. This
has mostly addressed the design of architectures, typically by adding parameters
and increasing model complexity. However, visual relation learning is a long-ta

iled problem, due to the combinatorial nature of joint reasoning about groups of objects. Increasing model complexity is, in general, ill-suited for long-tailed problems due to their tendency to overfit. In this paper, we explore an alternative hypothesis, denoted the Devil is in the Tails. Under this hypothesis, better performance is achieved by keeping the model simple but improving its ability to cope with long-tailed distributions. To test this hypothesis, we devise a new approach for training visual relationships models, which is inspired by state-of-the-art long-tailed recognition literature. This is based on an iterative decoupled training scheme, denoted Decoupled Training for Devil in the Tails (DT2). DT2 employs a novel sampling approach, Alternating Class-Balanced Sampling (ACBS), to capture the interplay between the long-tailed entity and predicate distributions of visual relations. Results show that, with an extremely simple architecture, DT2-ACBS significantly outperforms much more complex state-of-the-art methods on scene graph generation tasks. This suggests that the development of sophisticated models must be considered in tandem with the long-tailed nature of the problem.

****

T-SVDNet: Exploring High-Order Prototypical Correlations for Multi-Source Domain Adaptation

Ruihuang Li, Xu Jia, Jianzhong He, Shuaijun Chen, Qinghua Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9991-10000

Most existing domain adaptation methods focus on adaptation from only one source domain, however, in practice there are a number of relevant sources that could be leveraged to help improve performance on target domain. We propose a novel approach named T-SVDNet to address the task of Multi-source Domain Adaptation (MDA), which is featured by incorporating Tensor Singular Value Decomposition (T-SVD) into a neural network's training pipeline. Overall, high-order correlations among multiple domains are fully explored so as to better bridge the domain gap in this work. Specifically, we impose Tensor-Low-Rank (TLR) constraint on the tensor obtained by stacking up a group of prototypical similarity matrices, aiming at capturing consistent data structure across different domains. Furthermore, to avoid negative transfer brought by noisy source data, we propose a novel uncertainty-aware weighting strategy to adaptively assign weights to different source domains and samples based on the result of uncertainty estimation. Extensive experiments conducted on public benchmarks demonstrate the superiority of our model in addressing the task of MDA compared to state-of-the-art methods.

****

BuildingNet: Learning To Label 3D Buildings

Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, Evangelos Kalogerakis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10397-10407

We introduce BuildingNet: (a) a large-scale dataset of 3D building models whose exteriors are consistently labeled, and (b) a graph neural network that labels building meshes by analyzing spatial and structural relations of their geometric primitives. To create our dataset, we used crowdsourcing combined with expert guidance, resulting in 513K annotated mesh primitives, grouped into 292K semantic part components across 2K building models. The dataset covers several building categories, such as houses, churches, skyscrapers, town halls, libraries, and castles. We include a benchmark for evaluating mesh and point cloud labeling. Buildings have more challenging structural complexity compared to objects in existing benchmarks (e.g., ShapeNet, PartNet), thus, we hope that our dataset can nurture the development of algorithms that are able to cope with such large-scale geometric data for both vision and graphics tasks e.g., 3D semantic segmentation, part-based generative models, correspondences, texturing, and analysis of point cloud data acquired from real-world buildings. Finally, we show that our mesh-based graph neural network significantly improves performance over several baselines for labeling 3D meshes. Our project page www.buildingnet.org includes our dataset and code.

****
Student Customized Knowledge Distillation: Bridging the Gap Between Student and Teacher
Yichen Zhu, Yi Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5057-5066
Knowledge distillation (KD) transfers the dark knowledge from cumbersome networks (teacher) to lightweight (student) networks and expects the student to achieve more promising performance than training without the teacher's knowledge. However, a counter-intuitive argument is that better teachers do not make better students due to the capacity mismatch. To this end, we present a novel adaptive knowledge distillation method to complement traditional approaches. The proposed method, named as Student Customized Knowledge Distillation (SCKD), examines the capacity mismatch between teacher and student from the perspective of gradient similarity. We formulate the knowledge distillation as a multi-task learning problem so that the teacher transfers knowledge to the student only if the student can benefit from learning such knowledge. We validate our methods on multiple datasets with various teacher-student configurations on image classification, object detection, and semantic segmentation.
****
A Machine Teaching Framework for Scalable Recognition
Pei Wang, Nuno Vasconcelos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4945-4954
We consider the scalable recognition problem in the fine-grained expert domain where large-scale data collection is easy whereas annotation is difficult. Existing solutions are typically based on semi-supervised or self-supervised learning. We propose an alternative new framework, MEMORABLE, based on machine teaching and online crowdsourcing platforms. A small amount of data is first labeled by experts and then used to teach online annotators for the classes of interest, who finally label the entire dataset. Preliminary studies show that the accuracy of classifiers trained on the final dataset is a function of the accuracy of the student annotators. A new machine teaching algorithm, CMaxGrad, is then proposed to enhance this accuracy by introducing explanations in a state-of-the-art machine teaching algorithm. For this, CMaxGrad leverages counterfactual explanations, which take into account student predictions, thereby proving feedback that is student-specific, explicitly addresses the causes of student confusion, and adapts to the level of competence of the student. Experiments show that both MEMORABLE and CMaxGrad outperform existing solutions to their respective problems.
****
Divide and Conquer for Single-Frame Temporal Action Localization
Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13455-13464
Single-frame temporal action localization (STAL) aims to localize actions in untrimmed videos with only one timestamp annotation for each action instance. Existing methods adopt the one-stage framework but couple the counting goal and the localization goal. This paper proposes a novel two-stage framework for the STAL task with the spirit of divide and conquer. The instance counting stage leverages the location supervision to determine the number of action instances and divide a whole video into multiple video clips, so that each video clip contains only one complete action instance; and the location estimation stage leverages the category supervision to localize the action instance in each video clip. To efficiently represent the action instance in each video clip, we introduce the proposal-based representation, and design a novel differentiable mask generator to enable the end-to-end training supervised by category labels. On THUMOS14, GTEA, and BEOID datasets, our method outperforms state-of-the-art methods by 3.5%, 2.7%, 4.8% mAP on average. And extensive experiments verify the effectiveness of our method.
****
Real-World Video Super-Resolution: A Benchmark Dataset and a Decomposition Based Learning Scheme

Xi Yang, Wangmeng Xiang, Hui Zeng, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4781-4790
Video super-resolution (VSR) aims to improve the spatial resolution of low-resolution (LR) videos. Existing VSR methods are mostly trained and evaluated on synthetic datasets, where the LR videos are uniformly downsampled from their high-resolution (HR) counterparts by some simple operators (e.g., bicubic downsampling). Such simple synthetic degradation models, however, cannot well describe the complex degradation processes in real-world videos, and thus the trained VSR models become ineffective in real-world applications. As an attempt to bridge the gap, we build a real-world video super-resolution (RealVSR) dataset by capturing paired LR-HR video sequences using the multi-camera system of iPhone 11 Pro Max. Since the LR-HR video pairs are captured by two separate cameras, there are inevitably certain misalignment and luminance/color differences between them. To more robustly train the VSR model and recover more details from the LR inputs, we convert the LR-HR videos into YCbCr space and decompose the luminance channel into a Laplacian pyramid, and then apply different loss functions to different components. Experiments validate that VSR models trained on our RealVSR dataset demonstrate better visual quality than those trained on synthetic datasets under real-world settings. They also exhibit good generalization capability in cross-camera tests. The dataset and code can be found at https://github.com/IanYeung/RealVSR.
****

Towards High Fidelity Monocular Face Reconstruction With Rich Reflectance Using Self-Supervised Learning and Ray Tracing
Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, Louis Chevallier; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12819-12829
Robust face reconstruction from monocular image in general lighting conditions is challenging. Methods combining deep neural network encoders with differentiable rendering have opened up the path for very fast monocular reconstruction of geometry, lighting and reflectance. They can also be trained in self-supervised manner for increased robustness and better generalization. However, their differentiable rasterization based image formation models, as well as underlying scene parameterization, limit them to Lambertian face reflectance and to poor shape details. More recently, ray tracing was introduced for monocular face reconstruction within a classic optimization-based framework and enables state-of-the art results. However optimization-based approaches are inherently slow and lack robustness. In this paper, we build our work on the aforementioned approaches and propose a new method that greatly improves reconstruction quality and robustness in general scenes. We achieve this by combining a CNN encoder with a differentiable ray tracer, which enables us to base the reconstruction on much more advanced personalized diffuse and specular albedos, a more sophisticated illumination model and a plausible representation of self-shadows. This enables to take a big leap forward in reconstruction quality of shape, appearance and lighting even in scenes with difficult illumination. With consistent face attributes reconstruction, our method leads to practical applications such as relighting and self-shadows removal. Compared to state-of-the-art methods, our results show improved accuracy and validity of the approach.
****

Frequency Domain Image Translation: More Photo-Realistic, Better Identity-Preserving
Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, Gao Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13930-13940
Image-to-image translation has been revolutionized with GAN-based methods. However, existing methods lack the ability to preserve the identity of the source domain. As a result, synthesized images can often over-adapt to the reference domain, losing important structural characteristics and suffering from suboptimal visual quality. To solve these challenges, we propose a novel frequency domain image translation (FDIT) framework, exploiting frequency information for enhancing t

he image generation process. Our key idea is to decompose the image into low-fre
quency and high-frequency components, where the high-frequency feature captures
object structure akin to the identity. Our training objective facilitates the pr
eservation of frequency information in both pixel space and Fourier spectral spa
ce. We broadly evaluate FDIT across five large-scale datasets and multiple tasks
 including image translation and GAN inversion. Extensive experiments and ablati
ons show that FDIT effectively preserves the identity of the source image, and p
roduces photo-realistic images. FDIT establishes state-of-the-art performance, r
educing the average FID score by 5.6% compared to the previous best method.
****

ASCNet: Self-Supervised Video Representation Learning With Appearance-Speed Cons
istency
Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu,
 Mingkui Tan, Errui Ding; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2021, pp. 8096-8105
We study self-supervised video representation learning, which is a challenging t
ask due to 1) sufficient labels for supervision; 2) unstructured and noisy visua
l information. Existing methods mainly use contrastive loss with video clips as
the instances and learn visual representation by discriminating instances from e
ach other, but they need a careful treatment of negative pairs by either relying
 on large batch sizes, memory banks, extra modalities or customized mining strat
egies, which inevitably includes noisy data. In this paper, we observe that the
consistency between positive samples is the key to learn robust video representa
tion. Specifically, we propose two tasks to learn appearance and speed consisten
cy, respectively. The appearance consistency task aims to maximize the similarit
y between two clips of the same video with different playback speeds. The speed
consistency task aims to maximize the similarity between two clips with the same
 playback speed but different appearance information. We show that optimizing th
e two tasks jointly consistently improves the performance on downstream tasks, e
.g., action recognition and video retrieval. Remarkably, for action recognition
on the UCF-101 dataset, we achieve 90.8% accuracy without using any extra modali
ties or negative pairs for unsupervised pre-training, which outperforms the Imag
eNet supervised pre-trained model. Codes and models will be available.
****

Improving Generalization of Batch Whitening by Convolutional Unit Optimization
Yooshin Cho, Hanbyel Cho, Youngsoo Kim, Junmo Kim; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 5321-5329
Batch Whitening is a technique that accelerates and stabilizes training by trans
forming input features to have a zero mean (Centering) and a unit variance (Scal
ing), and by removing linear correlation between channels (Decorrelation). In co
mmonly used structures, which are empirically optimized with Batch Normalization
, the normalization layer appears between convolution and activation function. F
ollowing Batch Whitening studies have employed the same structure without furthe
r analysis; even Batch Whitening was analyzed on the premise that the input of a
 linear layer is whitened. To bridge the gap, we propose a new Convolutional Uni
t that in line with the theory, and our method generally improves the performanc
e of Batch Whitening. Moreover, we show the inefficacy of the original Convoluti
onal Unit by investigating rank and correlation of features. As our method is em
ployable off-the-shelf whitening modules, we use Iterative Normalization (IterNo
rm), the state-of-the-art whitening module, and obtain significantly improved pe
rformance on five image classification datasets: CIFAR-10, CIFAR-100, CUB-200-20
11, Stanford Dogs, and ImageNet. Notably, we verify that our method improves sta
bility and performance of whitening when using large learning rate, group size,
and iteration number.
****

Motion Guided Attention Fusion To Recognize Interactions From Videos
Tae Soo Kim, Jonathan Jones, Gregory D. Hager; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 13076-13086
We present a dual-pathway approach for recognizing fine-grained interactions fro
m videos. We build on the success of prior dual-stream approaches, but make a di

stinction between the static and dynamic representations of objects and their in teractions explicit by introducing separate motion and object detection pathways . Then, using our new Motion-Guided Attention Fusion module, we fuse the bottom- up features in the motion pathway with features captured from object detections to learn the temporal aspects of an action. We show that our approach can genera lize across appearance effectively and recognize actions where an actor interact s with previously unseen objects. We validate our approach using the composition al action recognition task from the Something-Something-v2 dataset where we outp erform existing state-of-the-art methods. We also show that our method can gener alize well to real world tasks by showing state-of-the-art performance on recogn izing humans assembling various IKEA furniture on the IKEA-ASM dataset.
****

Statistically Consistent Saliency Estimation
Shunyan Luo, Emre Barut, Fang Jin; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 745-753
The growing use of deep learning for a wide range of data problems has highlight ed the need to understand and diagnose these models appropriately, making deep l earning interpretation techniques an essential tool for data analysts. The numer ous model interpretation methods proposed in recent years are generally based on heuristics, with little or no theoretical guarantees. Here we present a statist ical framework for saliency estimation for black-box computer vision models. Our proposed model-agnostic estimation procedure, which is statistically consistent and capable of passing saliency checks, has polynomial-time computational effic iency since it only requires solving a linear program. An upper bound is establi shed on the number of model evaluations needed to recover regions of importance with high probability through our theoretical analysis. Furthermore, a new pertu rbation scheme is presented for the estimation of local gradients that is more e fficient than commonly used random perturbation schemes. The validity and excell ence of our new method are demonstrated experimentally using sensitivity analysi s on multiple datasets.
****

SLIDE: Single Image 3D Photography With Soft Layering and Depth-Aware Inpainting
Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T. Freeman, David Salesin, Brian Curless, Ce L iu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 12518-12527
Single image 3D photography enables viewers to view a still image from novel vie wpoints. Recent approaches combine monocular depth networks with inpainting netw orks to achieve compelling results. A drawback of these techniques is the use of hard depth layering, making them unable to model intricate appearance details s uch as thin hair-like structures. We present SLIDE, a modular and unified system for single image 3D photography that uses a simple yet effective soft layering strategy to better preserve appearance details in novel views. In addition, we p ropose a novel depth-aware training strategy for our inpainting module, better s uited for the 3D photography task. The resulting SLIDE approach is modular, enab ling the use of other components such as segmentation and matting for improved l ayering. At the same time, SLIDE uses an efficient layered depth formulation tha t only requires a single forward pass through the component networks to produce high quality 3D photos. Extensive experimental analysis on three view-synthesis datasets, in combination with user studies on in-the-wild image collections, dem onstrate superior performance of our technique in comparison to existing strong baselines while being conceptually much simpler. Project page: https://varunjamp ani.github.io/slide
****

Learning Spatio-Temporal Transformer for Visual Tracking
Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, Huchuan Lu; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10448-10457
In this paper, we present a new tracking architecture with an encoder-decoder tr ansformer as the key component. The encoder models the global spatio-temporal fe ature dependencies between target objects and search regions, while the decoder

learns a query embedding to predict the spatial positions of the target objects. Our method casts object tracking as a direct bounding box prediction problem, without using any proposals or predefined anchors. With the encoder-decoder transformer, the prediction of objects just uses a simple fully-convolutional network, which estimates the corners of objects directly. The whole method is end-to-end, does not need any postprocessing steps such as cosine window and bounding box smoothing, thus largely simplifying existing tracking pipelines. The proposed tracker achieves state-of-the-art performance on multiple challenging short-term and long-term benchmarks, while running at real-time speed, being 6x faster than Siam R-CNN. Code and models are open-sourced at https://github.com/researchmm/Stark.

****

From Contexts to Locality: Ultra-High Resolution Image Segmentation via Locality-Aware Contextual Correlation

Qi Li, Weixiang Yang, Wenxi Liu, Yuanlong Yu, Shengfeng He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7252-7261

Ultra-high resolution image segmentation has raised increasing interests in recent years due to its realistic applications. In this paper, we innovate the widely used high-resolution image segmentation pipeline, in which an ultra-high resolution image is partitioned into regular patches for local segmentation and then the local results are merged into a high-resolution semantic mask. In particular, we introduce a novel locality-aware contextual correlation based segmentation model to process local patches, where the relevance between local patch and its various contexts are jointly and complementarily utilized to handle the semantic regions with large variations. Additionally, we present a contextual semantics refinement network that associates the local segmentation result with its contextual semantics, and thus is endowed with the ability of reducing boundary artifacts and refining mask contours during the generation of final high-resolution mask. Furthermore, in comprehensive experiments, we demonstrate that our model outperforms other state-of-the-art methods in public benchmarks.

****

Channel-Wise Knowledge Distillation for Dense Prediction

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5311-5320

Knowledge distillation (KD) has been proven a simple and effective tool for training compact dense prediction models. Lightweight student networks are trained by extra supervision transferred from large teacher networks. Most previous KD variants for dense prediction tasks align the activation maps from the student and teacher network in the spatial domain, typically by normalizing the activation values on each spatial location and minimizing point-wise and/or pair-wise discrepancy. Different from the previous methods, here we propose to normalize the activation map of each channel to obtain a soft probability map. By simply minimizing the Kullback--Leibler (KL) divergence between the channel-wise probability map of the two networks, the distillation process pays more attention to the most salient regions of each channel, which are valuable for dense prediction tasks. We conduct experiments on a few dense prediction tasks, including semantic segmentation and object detection. Experiments demonstrate that our proposed method outperforms state-of-the-art distillation methods considerably, and can require less computational cost during training. In particular, we improve the RetinaNet detector (ResNet50backbone) by3.4%in mAP on the COCO dataset and spent (ResNet18 backbone) by5.81%in mIoU on the cityscapes dataset. Code is available at: https://git.io/Distiller.

****

Multi-View 3D Reconstruction With Transformers

Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, Rabab Ward; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5722-5731

Deep CNN-based methods have so far achieved the state of the art results in multi-view 3D object reconstruction. Despite the considerable progress, the two core

modules of these methods - view feature extraction and multi-view fusion, are usually investigated separately, and the relations among multiple input views are rarely explored. Inspired by the recent great success in Transformer models, we reformulate the multi-view 3D reconstruction as a sequence-to-sequence prediction problem and propose a framework named 3D Volume Transformer. Unlike previous CNN-based methods using a separate design, we unify the feature extraction and view fusion in a single Transformer network. A natural advantage of our design lies in the exploration of view-to-view relationships using self-attention among multiple unordered inputs. On ShapeNet - a large-scale 3D reconstruction benchmark, our method achieves a new state-of-the-art accuracy in multi-view reconstruction with fewer parameters (70% less) than CNN-based methods. Experimental results also suggest the strong scaling capability of our method. Our code will be made publicly available.

****

From General to Specific: Informative Scene Graph Generation via Balance Adjustment

Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, Jingkuan Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16383-16392

The scene graph generation (SGG) task aims to detect visual relationship triplets, i.e., subject, predicate, object, in an image, providing a structural vision layout for scene understanding. However, current models are stuck in common predicates, e.g., "on" and "at", rather than informative ones, e.g., "standing on" and "looking at", resulting in the loss of precise information and overall performance. If a model only uses "stone on road" rather than "blocking" to describe an image, it is easy to misunderstand the scene. We argue that this phenomenon is caused by two key imbalances between informative predicates and common ones, i.e., semantic space level imbalance and training sample level imbalance. To tackle this problem, we propose BA-SGG, a simple yet effective SGG framework based on balance adjustment but not the conventional distribution fitting. It integrates two components: Semantic Adjustment (SA) and Balanced Predicate Learning (BPL), respectively for adjusting these imbalances. Benefited from the model-agnostic process, our method is easily applied to the state-of-the-art SGG models and significantly improves the SGG performance. Our method achieves 14.3%, 8.0%, and 6.1% higher Mean Recall (mR) than that of the Transformer model at three scene graph generation sub-tasks on Visual Genome, respectively. Codes are publicly available.

****

Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering

Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, Zhaopeng Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13779-13788

Implicit neural rendering techniques have shown promising results for novel view synthesis. However, existing methods usually encode the entire scene as a whole, which is generally not aware of the object identity and limits the ability to the high-level editing tasks such as moving or adding furniture. In this paper, we present a novel neural scene rendering system, which learns an object-compositional neural radiance field and produces realistic rendering with editing capability for a clustered and real-world scene. Specifically, we design a novel two-pathway architecture, in which the scene branch encodes the scene geometry and appearance, and the object branch encodes each standalone object conditioned on learnable object activation codes. To survive the training in heavily cluttered scenes, we propose a scene-guided training strategy to solve the 3D space ambiguity in the occluded regions and learn sharp boundaries for each object. Extensive experiments demonstrate that our system not only achieves competitive performance for static scene novel-view synthesis, but also produces realistic rendering for object-level editing.

****

Practical Relative Order Attack in Deep Ranking

Mo Zhou, Le Wang, Zhenxing Niu, Qilin Zhang, Yinghui Xu, Nanning Zheng, Gang Hua

; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16413-16422
Recent studies unveil the vulnerabilities of deep ranking models, where an imperceptible perturbation can trigger dramatic changes in the ranking result. While previous attempts focus on manipulating absolute ranks of certain candidates, the possibility of adjusting their relative order remains under-explored. In this paper, we formulate a new adversarial attack against deep ranking systems, i.e., the Order Attack, which covertly alters the relative order among a selected set of candidates according to an attacker-specified permutation, with limited interference to other unrelated candidates. Specifically, it is formulated as a triplet-style loss imposing an inequality chain reflecting the specified permutation. However, direct optimization of such white-box objective is infeasible in a real-world attack scenario due to various black-box limitations. To cope with them, we propose a Short-range Ranking Correlation metric as a surrogate objective for black-box Order Attack to approximate the white-box method. The Order Attack is evaluated on the Fashion-MNIST and Stanford-Online-Products datasets under both white-box and black-box threat models. The black-box attack is also successfully implemented on a major e-commerce platform. Comprehensive experimental evaluations demonstrate the effectiveness of the proposed methods, revealing a new type of ranking model vulnerability.
****

Unsupervised Layered Image Decomposition Into Object Prototypes
Tom Monnier, Elliot Vincent, Jean Ponce, Mathieu Aubry; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8640-8650
We present an unsupervised learning framework for decomposing images into layers of automatically discovered object models. Contrary to recent approaches that model image layers with autoencoder networks, we represent them as explicit transformations of a small set of prototypical images. Our model has three main components: (i) a set of object prototypes in the form of learnable images with a transparency channel, which we refer to as sprites; (ii) differentiable parametric functions predicting occlusions and transformation parameters necessary to instantiate the sprites in a given image; (iii) a layered image formation model with occlusion for compositing these instances into complete images including background. By jointly learning the sprites and occlusion/transformation predictors to reconstruct images, our approach not only yields accurate layered image decompositions, but also identifies object categories and instance parameters. We first validate our approach by providing results on par with the state of the art on standard multi-object synthetic benchmarks (Tetrominoes, Multi-dSprites, CLEVR6). We then demonstrate the applicability of our model to real images in tasks that include clustering (SVHN, GTSRB), cosegmentation (Weizmann Horse) and object discovery from unfiltered social network images. To the best of our knowledge, our approach is the first layered image decomposition algorithm that learns an explicit and shared concept of object type, and is robust enough to be applied to real images.
****

Manifold Alignment for Semantically Aligned Style Transfer
Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14861-14869
Most existing style transfer methods follow the assumption that styles can be represented with global statistics (e.g., Gram matrices or covariance matrices), and thus address the problem by forcing the output and style images to have similar global statistics. An alternative is the assumption of local style patterns, where algorithms are designed to swap similar local features of content and style images. However, the limitation of these existing methods is that they neglect the semantic structure of the content image which may lead to corrupted content structure in the output. In this paper, we make a new assumption that image features from the same semantic region form a manifold and an image with multiple semantic regions follows a multi-manifold distribution. Based on this assumption, the style transfer problem is formulated as aligning two multi-manifold distrib

utions and a Manifold Alignment based Style Transfer (MAST) framework is propose d. The proposed framework allows semantically similar regions between the output and the style image share similar style patterns. Moreover, the proposed manifo ld alignment method is flexible to allow user editing or using semantic segmenta tion maps as guidance for style transfer. To allow the method to be applicable t o photorealistic style transfer, we propose a new adaptive weight skip connectio n network structure to preserve the content details. Extensive experiments verif y the effectiveness of the proposed framework for both artistic and photorealist ic style transfer. Code is available at https://github.com/NJUHuoJing/MAST.
****

Defending Against Universal Adversarial Patches by Clipping Feature Norms
Cheng Yu, Jiansheng Chen, Youze Xue, Yuyang Liu, Weitao Wan, Jiayu Bao, Huimin M a; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16434-16442
Physical-world adversarial attacks based on universal adversarial patches have b een proved to be able to mislead deep convolutional neural networks (CNNs), expo sing the vulnerability of real-world visual classification systems based on CNNs . In this paper, we empirically reveal and mathematically explain that the unive rsal adversarial patches usually lead to deep feature vectors with very large no rms in popular CNNs. Inspired by this, we propose a simple yet effective defendi ng approach using a new feature norm clipping (FNC) layer which is a differentia ble module that can be flexibly inserted in different CNNs to adaptively suppres s the generation of large norm deep feature vectors. FNC introduces no trainable parameter and only very low computational overhead. However, experiments on mul tiple datasets validate that it can effectively improve the robustness of differ ent CNNs towards white-box patch attacks while maintaining a satisfactory recogn ition accuracy for clean samples.
****

Q-Match: Iterative Shape Matching via Quantum Annealing
Marcel Seelbach Benkner, Zorah Lähner, Vladislav Golyanik, Christof Wunderlich, Christian Theobalt, Michael Moeller; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2021, pp. 7586-7596
Finding shape correspondences can be formulated as an NP-hard quadratic assignme nt problem (QAP) that becomes infeasible for shapes with high sampling density. A promising research direction is to tackle such quadratic optimization problems over binary variables with quantum annealing, which allows for some problems a more efficient search in the solution space. Unfortunately, enforcing the linear equality constraints in QAPs via a penalty significantly limits the success pro bability of such methods on currently available quantum hardware. To address thi s limitation, this paper proposes Q-Match, i.e., a new iterative quantum method for QAPs inspired by the alpha-expansion algorithm, which allows solving problem s of an order of magnitude larger than current quantum methods. It implicitly en forces the QAP constraints by updating the current estimates in a cyclic fashion . Further, Q-Match can be applied iteratively, on a subset of well-chosen corres pondences, allowing us to scale to real-world problems. Using the latest quantum annealer, the D-Wave Advantage, we evaluate the proposed method on a subset of QAPLIB as well as on isometric shape matching problems from the FAUST dataset.
****

Fast Convergence of DETR With Spatially Modulated Co-Attention
Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, Hongsheng Li; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 362 1-3630
The recently proposed Detection Transformer (DETR) model successfully applies Tr ansformer to objects detection and achieves comparable performance with two-stag e object detection frameworks, such as Faster-RCNN. However, DETR suffers from i ts slow convergence. Training DETR from scratch needs 500 epochs to achieve a hi gh accuracy. To accelerate its convergence, we propose a simple yet effective sc heme for improving the DETR framework, namely Spatially Modulated Co-Attention (SMCA) mechanism. The core idea of SMCA is to conduct location-aware co-attention in DETR by constraining co-attention responses to be high near initially estima

ted bounding box locations. Our proposed SMCA increases DETR's convergence speed by replacing the original co-attention mechanism in the decoder while keeping other operations in DETR unchanged. Furthermore, by integrating multi-head and scale-selection attention designs into SMCA, our fully-fledged SMCA can achieve better performance compared to DETR with a dilated convolution-based backbone (45.6 mAP at 108 epochs vs. 43.3 mAP at 500 epochs). We perform extensive ablation studies on COCO dataset to validate SMCA. Code is released at https://github.com/gaopengcuhk/SMCA-DETR.

****

Discovering Human Interactions With Large-Vocabulary Objects via Query and Multi-Scale Detection

Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, Yap-Peng Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13475-13484

In this work, we study the problem of human-object interaction (HOI) detection with large vocabulary object categories. Previous HOI studies are mainly conducted in the regime of limit object categories (e.g., 80 categories). Their solutions may face new difficulties in both object detection and interaction classification due to the increasing diversity of objects (e.g., 1000 categories). Different from previous methods, we formulate the HOI detection as a query problem. We propose a unified model to jointly discover the target objects and predict the corresponding interactions based on the human queries, thereby eliminating the need of using generic object detectors, extra steps to associate human-object instances, and multi-stream interaction recognition. This is achieved by a repurposed Transformer unit and a novel cascade detection over multi-scale feature maps. We observe that such a highly-coupled solution brings benefits for both object detection and interaction classification in a large vocabulary setting. To study the new challenges of the large vocabulary HOI detection, we assemble two datasets from the publicly available SWiG and 100 Days of Hands datasets. Experiments on these datasets validate that our proposed method can achieve a notable mAP improvement on HOI detection with a faster inference speed than existing one-stage HOI detectors.

****

T-AutoML: Automated Machine Learning for Lesion Segmentation Using Transformers in 3D Medical Imaging

Dong Yang, Andriy Myronenko, Xiaosong Wang, Ziyue Xu, Holger R. Roth, Daguang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3962-3974

Lesion segmentation in medical imaging has been an important topic in clinical research. Researchers have proposed various detection and segmentation algorithms to address this task. Recently, deep learning-based approaches have significantly improved the performance over conventional methods. However, most state-of-the-art deep learning methods require the manual design of multiple network components and training strategies. In this paper, we propose a new automated machine learning algorithm, T-AutoML, which not only searches for the best neural architecture, but also finds the best combination of hyper-parameters and data augmentation strategies simultaneously. The proposed method utilizes the modern transformer model, which is introduced to adapt to the dynamic length of the search space embedding and can significantly improve the ability of the search. We validate T-AutoML on several large-scale public lesion segmentation data-sets and achieve state-of-the-art performance.

****

CM-NAS: Cross-Modality Neural Architecture Search for Visible-Infrared Person Re-Identification

Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, Ran He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11823-11832

Visible-Infrared person re-identification (VI-ReID) aims to match cross-modality pedestrian images, breaking through the limitation of single-modality person ReID in dark environment. In order to mitigate the impact of large modality discre

pancy, existing works manually design various two-stream architectures to separa
tely learn modality-specific and modality-sharable representations. Such a manua
l design routine, however, highly depends on massive experiments and empirical p
ractice, which is time consuming and labor intensive. In this paper, we systemat
ically study the manually designed architectures, and identify that appropriatel
y separating Batch Normalization (BN) layers is the key to bring a great boost t
owards cross-modality matching. Based on this observation, the essential objecti
ve is to find the optimal separation scheme for each BN layer. To this end, we p
ropose a novel method, named Cross-Modality Neural Architecture Search (CM-NAS).
 It consists of a BN-oriented search space in which the standard optimization ca
n be fulfilled subject to the cross-modality task. Equipped with the searched ar
chitecture, our method outperforms state-of-the-art counterparts in both two ben
chmarks, improving the Rank-1/mAP by 6.70%/6.13% on SYSU-MM01 and by 12.17%/11.2
3% on RegDB. Code is released at https://github.com/JDAI-CV/CM-NAS.
****

Learning To Better Segment Objects From Unseen Classes With Unlabeled Videos
Yuming Du, Yang Xiao, Vincent Lepetit; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2021, pp. 3375-3384
The ability to localize and segment objects from unseen classes would open the d
oor to new applications, such as autonomous object learning in active vision. No
netheless, improving the performance on unseen classes requires additional train
ing data, while manually annotating the objects of the unseen classes can be lab
or-extensive and expensive. In this paper, we explore the use of unlabeled video
 sequences to automatically generate training data for objects of unseen classes
. It is in principle possible to apply existing video segmentation methods to un
labeled videos and automatically obtain object masks, which can then be used as
a training set even for classes with no manual labels available. However, our ex
periments show that these methods do not perform well enough for this purpose. W
e therefore introduce a Bayesian method that is specifically designed to automat
ically create such a training set: Our method starts from a set of object propos
als and relies on (non-realistic) analysis-by-synthesis to select the correct on
es by performing an efficient optimization over all the frames simultaneously. T
hrough extensive experiments, we show that our method can generate a high-qualit
y training set which significantly boosts the performance of segmenting objects
of unseen classes. We thus believe that our method could open the door for open-
world instance segmentation by exploiting abundant Internet videos.
****

SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervise
d Domain Adaptation
Viraj Prabhu, Shivam Khare, Deeksha Kartik, Judy Hoffman; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8558-8567
Many existing approaches for unsupervised domain adaptation (UDA) focus on adapt
ing under only data distribution shift and offer limited success under additiona
l cross-domain label distribution shift. Recent work based on self-training usin
g target pseudolabels has shown promise, but on challenging shifts pseudolabels
may be highly unreliable and using them for self-training may lead to error accu
mulation and domain misalignment. We propose Selective Entropy Optimization via
Committee Consistency (SENTRY), a UDA algorithm that judges the reliability of a
 target instance based on its predictive consistency under a committee of random
 image transformations. Our algorithm then selectively minimizes predictive entr
opy to increase confidence on highly consistent target instances, while maximizi
ng predictive entropy to reduce confidence on highly inconsistent ones. In combi
nation with pseudolabel-based approximate target class balancing, our approach l
eads to significant improvements over the state-of-the-art on 27/31 domain shift
s from standard UDA benchmarks as well as benchmarks designed to stress-test ada
ptation under label distribution shift.
****

Self-Supervised Domain Adaptation for Forgery Localization of JPEG Compressed Im
ages
Yuan Rao, Jiangqun Ni; Proceedings of the IEEE/CVF International Conference on C

omputer Vision (ICCV), 2021, pp. 15034-15043
With wide applications of image editing tools, forged images (splicing, copy-mov e, removal and etc.) have been becoming great public concerns. Although existing image forgery localization methods could achieve fairly good results on several public datasets, most of them perform poorly when the forged images are JPEG co mpressed as they are usually done in social networks. To tackle this issue, in t his paper, a self-supervised domain adaptation network, which is composed of a b ackbone network with Siamese architecture and a compression approximation networ k (ComNet), is proposed for JPEG-resistant image forgery localization. To improv e the performance against JPEG compression, ComNet is customized to approximate the JPEG compression operation through self-supervised learning, generating JPEG -agent images with general JPEG compression characteristics. The backbone networ k is then trained with domain adaptation strategy to localize the tampering boun dary and region, and alleviate the domain shift between uncompressed and JPEG-ag ent images. Extensive experimental results on several public datasets show that the proposed method outperforms or rivals to other state-of-the-art methods in i mage forgery localization, especially for JPEG compression with unknown QFs.
****

Unsupervised Point Cloud Object Co-Segmentation by Co-Contrastive Learning and M utual Attention Sampling
Cheng-Kun Yang, Yung-Yu Chuang, Yen-Yu Lin; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2021, pp. 7335-7344
This paper presents a new task, point cloud object co-segmentation, aiming to se gment the common 3D objects in a set of point clouds. We formulate this task as an object point sampling problem, and develop two techniques, the mutual attenti on module and co-contrastive learning, to enable it. The proposed method employs two point samplers based on deep neural networks, the object sampler and the ba ckground sampler. The former targets at sampling points of common objects while the latter focuses on the rest. The mutual attention module explores point-wise correlation across point clouds. It is embedded in both samplers and can identif y points with strong cross-cloud correlation from the rest. After extracting fea tures for points selected by the two samplers, we optimize the networks by devel oping the co-contrastive loss, which minimizes feature discrepancy of the estima ted object points while maximizing feature separation between the estimated obje ct and background points. Our method works on point clouds of an arbitrary objec t class. It is end-to-end trainable and does not need point-level annotations. I t is evaluated on the ScanObjectNN and S3DIS datasets and achieves promising res ults.
****

Meta Pairwise Relationship Distillation for Unsupervised Person Re-Identificatio n
Haoxuanye Ji, Le Wang, Sanping Zhou, Wei Tang, Nanning Zheng, Gang Hua; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp . 3661-3670
Unsupervised person re-identification (Re-ID) remains challenging due to the lac k of ground-truth labels. Existing methods often rely on estimated pseudo labels via iterative clustering and classification, and they are unfortunately highly susceptible to performance penalties incurred by the inaccurate estimated number of clusters. Alternatively, we propose the Meta Pairwise Relationship Distillat ion (MPRD) method to estimate the pseudo labels of sample pairs for unsupervised person Re-ID. Specifically, it consists of a Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN), in which the GCN estimates the pseudo la bels of sample pairs based on the current features extracted by CNN, and the CNN learns better features by involving high-fidelity positive and negative sample pairs imposed by GCN. To achieve this goal, a small amount of labeled samples ar e used to guide GCN training, which can distill meta knowledge to judge the diff erence in the neighborhood structure between positive and negative sample pairs. Extensive experiments on Market-1501, DukeMTMC-reID and MSMT17 datasets show th at our method outperforms the state-of-the-art approaches.
****

Relational Embedding for Few-Shot Classification
Dahyun Kang, Heeseung Kwon, Juhong Min, Minsu Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8822-8833
We propose to address the problem of few-shot classification by meta-learning "what to observe" and "where to attend" in a relational perspective. Our method leverages relational patterns within and between images via self-correlational representation (SCR) and cross-correlational attention (CCA). Within each image, the SCR module transforms a base feature map into a self-correlation tensor and learns to extract structural patterns from the tensor. Between the images, the CCA module computes cross-correlation between two image representations and learns to produce co-attention between them. Our Relational Embedding Network (RENet) combines the two relational modules to learn relational embedding in an end-to-end manner. In experimental evaluation, it achieves consistent improvements over state-of-the-art methods on four widely used few-shot classification benchmarks of miniImageNet, tieredImageNet, CUB-200-2011, and CIFAR-FS.
****
Globally Optimal and Efficient Manhattan Frame Estimation by Delimiting Rotation Search Space
Wuwei Ge, Yu Song, Baichao Zhang, Zehua Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15213-15221
A typical man-made structure can be abstracted as the Manhattan world assumption, in which notion is further represented as a Manhattan Frame (MF) defined by three orthogonal axes. The problem of MF estimation can be formulated as the solution of the rotation between the MF and the camera frame (called the "MF rotation"). However, the whole rotation space is quite redundant for solving the MF rotation, which is one of the main factors that disturb the computational efficiency of those methods associated with a rotation space search. This paper proves that the volume of the space that just contains all MF rotations (called the "MFR space") is only 1 / 24 of that of the whole rotation space, and then an exact MFR space is delimited from the rotation space. Searching in the delimited MFR space, the MF estimation solved by a branch-and-bound (BnB) framework guarantees stability and efficiency simultaneously. Furthermore, the general rotation problems associated with a rotation space search are solved more efficiently. Experiments on synthetic and real datasets have successfully confirmed the validity of our approach.
****
Robustness Certification for Point Cloud Models
Tobias Lorenz, Anian Ruoss, Mislav Balunovi■, Gagandeep Singh, Martin Vechev; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7608-7618
The use of deep 3D point cloud models in safety-critical applications, such as autonomous driving, dictates the need to certify the robustness of these models to real-world transformations. This is technically challenging, as it requires a scalable verifier tailored to point cloud models that handles a wide range of semantic 3D transformations. In this work, we address this challenge and introduce 3DCertify, the first verifier able to certify the robustness of point cloud models. 3DCertify is based on two key insights: (i) a generic relaxation based on first-order Taylor approximations, applicable to any differentiable transformation, and (ii) a precise relaxation for global feature pooling, which is more complex than pointwise activations (e.g., ReLU or sigmoid) but commonly employed in point cloud models. We demonstrate the effectiveness of 3DCertify by performing an extensive evaluation on a wide range of 3D transformations (e.g., rotation, twisting) for both classification and part segmentation tasks. For example, we can certify robustness against rotations by +-60deg for 95.7% of point clouds, and our max pool relaxation increases certification by up to 15.6%.
****
Square Root Marginalization for Sliding-Window Bundle Adjustment
Nikolaus Demmel, David Schubert, Christiane Sommer, Daniel Cremers, Vladyslav Usenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13260-13268

In this paper we propose a novel square root sliding-window bundle adjustment suitable for real-time odometry applications. The square root formulation pervades three major aspects of our optimization-based sliding-window estimator: for bundle adjustment we eliminate landmark variables with nullspace projection; to store the marginalization prior we employ a matrix square root of the Hessian; and when marginalizing old poses we avoid forming normal equations and update the square root prior directly with a specialized QR decomposition. We show that the proposed square root marginalization is algebraically equivalent to the conventional use of Schur complement (SC) on the Hessian. Moreover, it elegantly deals with rank-deficient Jacobians producing a prior equivalent to SC with Moore--Penrose inverse. Our evaluation of visual and visual-inertial odometry on real-world datasets demonstrates that the proposed estimator is 36% faster than the baseline. It furthermore shows that in single precision, conventional Hessian-based marginalization leads to numeric failures and reduced accuracy. We analyse numeric properties of the marginalization prior to explain why our square root form does not suffer from the same effect and therefore entails superior performance.
****

OpenForensics: Large-Scale Challenging Dataset for Multi-Face Forgery Detection and Segmentation In-the-Wild

Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10117-10127

The proliferation of deepfake media is raising concerns among the public and relevant authorities. It has become essential to develop countermeasures against forged faces in social media. This paper presents a comprehensive study on two new countermeasure tasks: multi-face forgery detection and segmentation in-the-wild. Localizing forged faces among multiple human faces in unrestricted natural scenes is far more challenging than the traditional deepfake recognition task. To promote these new tasks, we have created the first large-scale dataset posing a high level of challenges that is designed with face-wise rich annotations explicitly for face forgery detection and segmentation, namely OpenForensics. With its rich annotations, our OpenForensics dataset has great potentials for research in both deepfake prevention and general human face detection. We have also developed a suite of benchmarks for these tasks by conducting an extensive evaluation of state-of-the-art instance detection and segmentation methods on our newly constructed dataset in various scenarios.
****

Hierarchical Conditional Flow: A Unified Framework for Image Super-Resolution and Image Rescaling

Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4076-4085

Normalizing flows have recently demonstrated promising results for low-level vision tasks. For image super-resolution (SR), it learns to predict diverse photo-realistic high-resolution (HR) images from the low-resolution (LR) image rather than learning a deterministic mapping. For image rescaling, it achieves high accuracy by jointly modelling the downscaling and upscaling processes. While existing approaches employ specialized techniques for these two tasks, we set out to unify them in a single formulation. In this paper, we propose the hierarchical conditional flow (HCFlow) as a unified framework for image SR and image rescaling. More specifically, HCFlow learns a bijective mapping between HR and LR image pairs by modelling the distribution of the LR image and the rest high-frequency component simultaneously. In particular, the high-frequency component is conditional on the LR image in a hierarchical manner. To further enhance the performance, other losses such as perceptual loss and GAN loss are combined with the commonly used negative log-likelihood loss in training. Extensive experiments on general image SR, face image SR and image rescaling have demonstrated that the proposed HCFlow achieves state-of-the-art performance in terms of both quantitative metrics and visual quality.
****

Deep Symmetric Network for Underexposed Image Enhancement With Recurrent Attentional Learning
Lin Zhao, Shao-Ping Lu, Tao Chen, Zhenglu Yang, Ariel Shamir; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12075-12084
Underexposed image enhancement is of importance in many research domains. In this paper, we take this problem as image feature transformation between the underexposed image and its paired enhanced version, and we propose a deep symmetric network for the issue. Our symmetric network adapts invertible neural networks (INN) for bidirectional feature learning between images, and to ensure the mutual propagation invertible we specifically construct two pairs of encoder-decoder with the same pretrained parameters. This invertible mechanism with bidirectional feature transformations enable us to both avoid colour bias and recover the content effectively for image enhancement. In addition, we propose a new recurrent residual-attention module (RRAM), where the recurrent learning network is designed to gradually perform the desired colour adjustments. Ablation experiments are executed to show the role of each component of our new architecture. We conduct a large number of experiments on two datasets to demonstrate that our method achieves the state-of-the-art effect in underexposed image enhancement. Code is available at https://www.shaopinglu.net/proj-iccv21/ImageEnhancement.html
****

Syncretic Modality Collaborative Learning for Visible Infrared Person Re-Identification
Ziyu Wei, Xi Yang, Nannan Wang, Xinbo Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 225-234
Visible infrared person re-identification (VI-REID) aims to match pedestrian images between the daytime visible and nighttime infrared camera views. The large cross-modality discrepancies have become the bottleneck which limits the performance of VI-REID. Existing methods mainly focus on capturing cross-modality sharable representations by learning an identity classifier. However, the heterogeneous pedestrian images taken by different spectrum cameras differ significantly in image styles, resulting in inferior discriminability of feature representations. To alleviate the above problem, this paper explores the correlation between two modalities and proposes a novel syncretic modality collaborative learning (SMCL) model to bridge the cross-modality gap. A new modality that incorporates features of heterogeneous images is constructed automatically to steer the generation of modality-invariant representations. Challenge enhanced homogeneity learning (CEHL) and auxiliary distributional similarity learning (ADSL) are integrated to project heterogeneous features on a unified space and enlarge the inter-class disparity, thus strengthening the discriminative power. Extensive experiments on two cross-modality benchmarks demonstrate the effectiveness and superiority of the proposed method. Especially, on SYSU-MM01 dataset, our SMCL model achieves 67.39% rank-1 accuracy and 61.78% mAP, surpassing the cutting-edge works by a large margin.
****

Estimating Egocentric 3D Human Pose in Global Space
Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Christian Theobalt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11500-11509
Egocentric 3D human pose estimation using a single fisheye camera has become popular recently as it allows capturing a wide range of daily activities in unconstrained environments, which is difficult for traditional outside-in motion capture with external cameras. However, existing methods have several limitations. A prominent problem is that the estimated poses lie in the local coordinate system of the fisheye camera, rather than in the world coordinate system, which is restrictive for many applications. Furthermore, these methods suffer from limited accuracy and temporal instability due to ambiguities caused by the monocular setup and the severe occlusion in a strongly distorted egocentric perspective. To tackle these limitations, we present a new method for egocentric global 3D body pose estimation using a single head-mounted fisheye camera. To achieve accurate and

temporally stable global poses, a spatio-temporal optimization is performed ove
r a sequence of frames by minimizing heatmap reprojection errors and enforcing l
ocal and global body motion priors learned from a mocap dataset. Experimental re
sults show that our approach outperforms state-of-the-art methods both quantitat
ively and qualitatively.
****

Re-Aging GAN: Toward Personalized Face Age Transformation
Farkhod Makhmudkhujaev, Sungeun Hong, In Kyu Park; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 3908-3917
Face age transformation aims to synthesize past or future face images by reflect
ing the age factor on given faces. Ideally, this task should synthesize natural-
looking faces across various age groups while maintaining identity. However, mos
t of the existing work has focused on only one of these or is difficult to train
while unnatural artifacts still appear. In this work, we propose Re-Aging GAN (
RAGAN), a novel single framework considering all the critical factors in age tra
nsformation. Our framework achieves state-of-the-art personalized face age trans
formation by compelling the input identity to perform the self-guidance of the g
eneration process. Specifically, RAGAN can learn the personalized age features b
y using high-order interactions between given identity and target age. Learned p
ersonalized age features are identity information that is recalibrated according
to the target age. Hence, such features encompass identity and target age infor
mation that provides important clues on how an input identity should be at a cer
tain age. Experimental result shows the lowest FID and KID scores and the highes
t age recognition accuracy compared to previous methods. The proposed method als
o demonstrates the visual superiority with fewer artifacts, identity preservatio
n, and natural transformation across various age groups.
****

SeLFVi: Self-Supervised Light-Field Video Reconstruction From Stereo Video
Prasan Shedligeri, Florian Schiffers, Sushobhan Ghosh, Oliver Cossairt, Kaushik
Mitra; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2021, pp. 2491-2501
Light-field (LF) imaging is appealing to the mobile devices market because of it
s capability for intuitive post-capture processing. Acquiring LF data with high
angular, spatial and temporal resolution poses significant challenges, especiall
y with space constraints preventing bulky optics. At the same time, stereo video
capture, now available on many consumer devices, can be interpreted as a sparse
LF-capture. We explore the application of small baseline stereo videos for reco
nstructing high fidelity LF videos. We propose a self-supervised learning-based
algorithm for LF video reconstruction from stereo video. The self-supervised LF
video reconstruction is guided via the geometric information from the individual
stereo pairs and the temporal information from the video sequence. LF estimatio
n is further regularized by a low-rank constraint based on layered LF displays.
The proposed self-supervised algorithm facilitates advantages such as post-train
ing fine-tuning on test sequences and variable angular view interpolation and ex
trapolation. Quantitatively the LF videos show higher fidelity than previously p
roposed unsupervised approaches for LF reconstruction. We demonstrate our result
s via LF videos generated from stereo videos acquired from commercially availabl
e stereoscopic cameras. Finally, we demonstrate that our reconstructed LF videos
allow applications such as post-capture focus control and RoI-based focus track
ing for videos.
****

Cross-Encoder for Unsupervised Gaze Representation Learning
Yunjia Sun, Jiabei Zeng, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2021, pp. 3702-3711
In order to train 3D gaze estimators without too many annotations, we propose an
unsupervised learning framework, Cross-Encoder, to leverage the unlabeled data
to learn suitable representation for gaze estimation. To address the issue that
the feature of gaze is always intertwined with the appearance of the eye, Cross-
Encoder disentangles the features using a latent-code-swapping mechanism on eye-
consistent image pairs and gaze-similar ones. Specifically, each image is encode

d as a gaze feature and an eye feature. Cross-Encoder is trained to reconstruct each image in the eye-consistent pair according to its gaze feature and the other's eye feature, but to reconstruct each image in the gaze-similar pair according to its eye feature and the other's gaze feature. Experimental results show the validity of our work. First, using the Cross-Encoder-learned gaze representation, the gaze estimator trained with very few samples outperforms the ones using other unsupervised learning methods, under both within-dataset and cross-dataset protocol. Second, ResNet18 pretrained by Cross-Encoder is competitive with state-of-the-art gaze estimation methods. Third, ablation study shows that Cross-Encoder disentangles the gaze feature and eye feature.
****

Towards Discriminative Representation Learning for Unsupervised Person Re-Identification
Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, Shengjin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8526-8536
In this work, we address the problem of unsupervised domain adaptation for person re-ID where annotations are available for the source domain but not for target. Previous methods typically follow a two-stage optimization pipeline, where the network is first pre-trained on source and then fine-tuned on target with pseudo labels created by feature clustering. Such methods sustain two main limitations. (1) The label noise may hinder the learning of discriminative features for recognizing target classes. (2) The domain gap may hinder knowledge transferring from source to target. We propose three types of technical schemes to alleviate these issues. First, we propose a cluster-wise contrastive learning algorithm (CCL) by iterative optimization of feature learning and cluster refinery to learn noise-tolerant representations in the unsupervised manner. Second, we adopt a progressive domain adaptation (PDA) strategy to gradually mitigate the domain gap between source and target data. Third, we propose Fourier augmentation (FA) for further maximizing the class separability of re-ID models by imposing extra constraints in the Fourier space. We observe that these proposed schemes are capable of facilitating the learning of discriminative feature representations. Experiments demonstrate that our method consistently achieves notable improvements over the state-of-the-art unsupervised re-ID methods on multiple benchmarks, e.g., surpassing MMT largely by 8.1%, 9.9%, 11.4% and 11.1% mAP on the Market-to-Duke, Duke-to-Market, Market-to-MSMT and Duke-to-MSMT tasks, respectively.
****

Event-Intensity Stereo: Estimating Depth by the Best of Both Worlds
Mohammad Mostafavi, Kuk-Jin Yoon, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4258-4267
Event cameras can report scene movements as an asynchronous stream of data called the events. Unlike traditional cameras, event cameras have very low latency (microseconds vs milliseconds) very high dynamic range (140dB vs 60 dB), and low power consumption, as they report changes of a scene and not a complete frame. As they re-port per pixel feature-like events and not the whole intensity frame they are immune to motion blur. However, event cameras require movement between the scene and camera to fire events ,i.e., they have no output when the scene is relatively static. Traditional cameras, however, report the whole frame of pixels at once in fixed intervals but have lower dynamic range and are prone to motion blur in case of rapid movements. We get the best from both worlds and use events and intensity images together in our complementary design and estimate dense disparity from this combination. The proposed end-to-end design combines events and images in a sequential manner and correlates them to esti-mate dense depth values. Our various experimental settings in real-world and simulated scenarios exploit the superiority of our method in predicting accurate depth values with fine details. We further extend our method to extreme cases of missing the left or right event or stereo pair and also investigate stereo depth estimation with inconsistent dynamic ranges or event thresholds on the left and right pairs
****

When Do GANs Replicate? On the Choice of Dataset Size

Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, Aleix M. Martinez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6701-6710

Do GANs replicate training images? Previous studies have shown that GANs do not seem to replicate training data without significant change in the training procedure. This leads to a series of research on the exact condition needed for GANs to overfit to the training data. Although a number of factors has been theoretically or empirically identified, the effect of dataset size and complexity on GANs replication is still unknown. With empirical evidence from BigGAN and StyleGAN2, on datasets CelebA, Flower and LSUN-bedroom, we show that dataset size and its complexity play an important role in GANs replication and perceptual quality of the generated images. We further quantify this relationship, discovering that replication percentage decays exponentially with respect to dataset size and complexity, with a shared decaying factor across GAN-dataset combinations. Meanwhile, the perceptual image quality follows a U-shape trend w.r.t dataset size. This finding leads to a practical tool for one-shot estimation on minimal dataset size to prevent GAN replication which can be used to guide datasets construction and selection.

****

Just One Moment: Structural Vulnerability of Deep Action Recognition Against One Frame Attack

Jaehui Hwang, Jun-Hyuk Kim, Jun-Ho Choi, Jong-Seok Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7668-7676

The video-based action recognition task has been extensively studied in recent years. In this paper, we study the structural vulnerability of deep learning-based action recognition models against the adversarial attack using the one frame attack that adds an inconspicuous perturbation to only a single frame of a given video clip. Our analysis shows that the models are highly vulnerable against the one frame attack due to their structural properties. Experiments demonstrate high fooling rates and inconspicuous characteristics of the attack. Furthermore, we show that strong universal one frame perturbations can be obtained under various scenarios. Our work raises the serious issue of adversarial vulnerability of the state-of-the-art action recognition models in various perspectives.

****

Big Self-Supervised Models Advance Medical Image Classification

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, Mohammad Norouzi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3478-3488

Self-supervised pretraining followed by supervised fine-tuning has seen success in image recognition, especially when labeled examples are scarce, but has received limited attention in medical image analysis. This paper studies the effectiveness of self-supervised learning as a pretraining strategy for medical image classification. We conduct experiments on two distinct tasks: dermatology condition classification from digital camera images and multi-label chest X-ray classification, and demonstrate that self-supervised learning on ImageNet, followed by additional self-supervised learning on unlabeled domain-specific medical images significantly improves the accuracy of medical image classifiers.We introduce a novel Multi-Instance Contrastive Learning (MICLe) method that uses multiple images of the underlying pathology per patient case, when available, to construct more informative positive pairs for self-supervised learning. Combining our contributions, we achieve an improvement of 6.7% in top-1 accuracy and an improvement of 1.1% in mean AUC on dermatology and chest X-ray classification respectively, outperforming strong supervised baselines pretrained on ImageNet. In addition, we show that big self-supervised models are robust to distribution shift and can learn efficiently with a small number of labeled medical images.

****

Scene Context-Aware Salient Object Detection

Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,

pp. 4156-4166
Salient object detection identifies objects in an image that grab visual attenti
on. Although contextual features are considered in recent literature, they often
 fail in real-world complex scenarios. We observe that this is mainly due to two
 issues: First, most existing datasets consist of simple foregrounds and backgro
unds that hardly represent real-life scenarios. Second, current methods only lea
rn contextual features of salient objects, which are insufficient to model high-
level semantics for saliency reasoning in complex scenes. To address these probl
ems, we first construct a new large-scale dataset with complex scenes in this pa
per. We then propose a context-aware learning approach to explicitly exploit the
 semantic scene contexts. Specifically, two modules are proposed to achieve the
goal: 1) a Semantic Scene Context Refinement module to enhance contextual featur
es learned from salient objects with scene context, and 2) a Contextual Instance
 Transformer to learn contextual relations between objects and scene context. To
 our knowledge, such high-level semantic contextual information of image scenes
is under-explored for saliency detection in the literature. Extensive experiment
s demonstrate that the proposed approach outperforms state-of-the-art techniques
 in complex scenarios for saliency detection, and transfers well to other existi
ng datasets. The code and dataset are available at https://github.com/SirisAvish
ek/Scene_Context_Aware_Saliency.
****

Learning Frequency-Aware Dynamic Network for Efficient Super-Resolution
Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, Yunhe Wang; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
4308-4317
Deep learning based methods, especially convolutional neural networks (CNNs) hav
e been successfully applied in the field of single image super-resolution (SISR)
. To obtain better fidelity and visual quality, most of existing networks are of
 heavy design with massive computation. However, the computation resources of mo
dern mobile devices are limited, which cannot easily support the expensive cost.
 To this end, this paper explores a novel frequency-aware dynamic network for di
viding the input into multiple parts according to its coefficients in the discre
te cosine transform (DCT) domain. In practice, the high-frequency part will be p
rocessed using expensive operations and the lower-frequency part is assigned wit
h cheap operations to relieve the computation burden. Since pixels or image patc
hes belong to low-frequency areas contain relatively few textural details, this
dynamic network will not affect the quality of resulting super-resolution images
. In addition, we embed predictors into the proposed dynamic network to end-to-e
nd fine-tune the handcrafted frequency-aware masks. Extensive experiments conduc
ted on benchmark SISR models and datasets show that the frequency-aware dynamic
network can be employed for various SISR neural architectures to obtain the bett
er tradeoff between visual quality and computational complexity. For instance, w
e can reduce the FLOPs of SR models by approximate 50% while preserving the stat
e-of-the-art SISR performance.
****

Road Anomaly Detection by Partial Image Reconstruction With Segmentation Couplin
g
Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino,
Jiri Matas; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2021, pp. 15651-15660
We present a novel approach to the detection of unknown objects in the context o
f autonomous driving. The problem is formulated as anomaly detection, since we a
ssume that the unknown stuff or object appearance cannot be learned. To that end
, we propose a reconstruction module that can be used with many existing semanti
c segmentation networks, and that is trained to recognize and reconstruct road (
drivable) surface from a small bottleneck. We postulate that poor reconstruction
 of the road surface is due to areas that are outside of the training distributi
on, which is a strong indicator of an anomaly. The road structural similarity er
ror is coupled with the semantic segmentation to incorporate information from kn
own classes and produce final per-pixel anomaly scores. The proposed JSR-Net was

evaluated on four datasets, Lost-and-found, Road Anomaly, Road Obstacles, and FishyScapes, achieving state-of-art performance on all, reducing the false positives significantly, while typically having the highest average precision for wide range of operation points.

****

BossNAS: Exploring Hybrid CNN-Transformers With Block-Wisely Self-Supervised Neural Architecture Search

Changlin Li, Tao Tang, Guangrun Wang, Jiefeng Peng, Bing Wang, Xiaodan Liang, Xiaojun Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12281-12291

A myriad of recent breakthroughs in hand-crafted neural architectures for visual recognition have highlighted the urgent need to explore hybrid architectures consisting of diversified building blocks. Meanwhile, neural architecture search methods are surging with an expectation to reduce human efforts. However, whether NAS methods can efficiently and effectively handle diversified search spaces with disparate candidates (e.g. CNNs and transformers) is still an open question. In this work, we present Block-wisely Self-supervised Neural Architecture Search (BossNAS), an unsupervised NAS method that addresses the problem of inaccurate architecture rating caused by large weight-sharing space and biased supervision in previous methods. More specifically, we factorize the search space into blocks and utilize a novel self-supervised training scheme, named ensemble bootstrapping, to train each block separately before searching them as a whole towards the population center. Additionally, we present HyTra search space, a fabric-like hybrid CNN-transformer search space with searchable down-sampling positions. On this challenging search space, our searched model, BossNet-T, achieves up to 82.5% accuracy on ImageNet, surpassing EfficientNet by 2.4% with comparable compute time. Moreover, our method achieves superior architecture rating accuracy with 0.78 and 0.76 Spearman correlation on the canonical MBConv search space with ImageNet and on NATS-Bench size search space with CIFAR-100, respectively, surpassing state-of-the-art NAS methods.

****

H2O: Two Hands Manipulating Objects for First Person Interaction Recognition

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10138-10148

We present a comprehensive framework for egocentric interaction recognition using markerless 3D annotations of two hands manipulating objects. To this end, we propose a method to create a unified dataset for egocentric 3D interaction recognition. Our method produces annotations of the 3D pose of two hands and the 6D pose of the manipulated objects, along with their interaction labels for each frame. Our dataset, called H2O (2 Hands and Objects), provides synchronized multi-view RGB-D images, interaction labels, object classes, ground-truth 3D poses for left & right hands, 6D object poses, ground-truth camera poses, object meshes and scene point clouds. To the best of our knowledge, this is the first benchmark that enables the study of first-person actions with the use of the pose of both left and right hands manipulating objects and presents an unprecedented level of detail for egocentric 3D interaction recognition. We further propose the method to predict interaction classes by estimating the 3D pose of two hands and the 6D pose of the manipulated objects, jointly from RGB images. Our method models both inter- and intra-dependencies between both hands and objects by learning the topology of a graph convolutional network that predicts interactions. We show that our method facilitated by this dataset establishes a strong baseline for joint hand-object pose estimation and achieves state-of-the-art accuracy for first person interaction recognition.

****

Residual Attention: A Simple but Effective Method for Multi-Label Recognition

Ke Zhu, Jianxin Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 184-193

Multi-label image recognition is a challenging computer vision task of practical use. Progresses in this area, however, are often characterized by complicated m

ethods, heavy computations, and lack of intuitive explanations. To effectively c
apture different spatial regions occupied by objects from different categories,
we propose an embarrassingly simple module, named class-specific residual attent
ion (CSRA). CSRA generates class-specific features for every category by proposi
ng a simple spatial attention score, and then combines it with the class-agnosti
c average pooling feature. CSRA achieves state-of-the-art results on multilabel
recognition, and at the same time is much simpler than them. Furthermore, with o
nly 4 lines of code, CSRA also leads to consistent improvement across many diver
se pretrained models and datasets without any extra training. CSRA is both easy
to implement and light in computations, which also enjoys intuitive explanations
 and visualizations.
****

TransferI2I: Transfer Learning for Image-to-Image Translation From Small Dataset
s

Yaxing Wang, Héctor Laria, Joost van de Weijer, Laura Lopez-Fuentes, Bogdan Radu
canu; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2021, pp. 14010-14019

Image-to-image (I2I) translation has matured in recent years and is able to gene
rate high-quality realistic images. However, despite current success, it still f
aces important challenges when applied to small domains. Existing methods use tr
ansfer learning for I2I translation, but they still require the learning of mill
ions of parameters from scratch. This drawback severely limits its application o
n small domains. In this paper, we propose a new transfer learning for I2I trans
lation (TransferI2I). We decouple our learning process into the image generation
 step and the I2I translation step. In the first step we propose two novel techn
iques: source-target initialization and self-initialization of the adaptor layer
. The former finetunes the pretrained generative model (e.g., StyleGAN) on sourc
e and target data. The latter allows to initialize all non-pretrained network pa
rameters without the need of any data. These techniques provide a better initial
ization for the I2I translation. Second step performs the actual I2I translation
 using the learned weights in the first step. In addition, we introduce an auxil
iary GAN that further facilitates the training of deep I2I systems even from sma
ll datasets. In extensive experiments on three datasets, (Animal faces, Birds, a
nd Foods), we show that we outperform existing methods and that mFID improves on
 several datasets with over 25 points.
****

On Generating Transferable Targeted Perturbations

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Fatih Porikli;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2021, pp. 7708-7717

While the untargeted black-box transferability of adversarial perturbations has
been extensively studied before, changing an unseen model's decisions to a speci
fic `targeted' class remains a challenging feat. In this paper, we propose a new
 generative approach for highly transferable targeted perturbations (\ours). We
note that the existing methods are less suitable for this task due to their reli
ance on class-boundary information that changes from one model to another, thus
reducing transferability. In contrast, our approach matches perturbed image `dis
tribution' with that of the target class, leading to high targeted transferabili
ty rates. To this end, we propose a new objective function that not only aligns
the global distributions of source and target images, but also matches the local
 neighbourhood structure between the two domains. Based on the proposed objectiv
e, we train a generator function that can adaptively synthesize perturbations sp
ecific to a given input. Our generative approach is independent of the source or
 target domain labels, while consistently performs well against state-of-the-art
 methods on a wide range of attack settings. As an example, we achieve 32.63% ta
rget transferability from (an adversarially weak) VGG19_ BN to (a strong) WideR
esNet on ImageNet val. set, which is 4xhigher than the previous best generative
attack and 16xbetter than instance-specific iterative attack.
****

SynFace: Face Recognition With Synthetic Data

Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10880-10890

With the recent success of deep neural networks, remarkable progress has been achieved on face recognition. However, collecting large-scale real-world training data for face recognition has turned out to be challenging, especially due to the label noise and privacy issues. Meanwhile, existing face recognition datasets are usually collected from web images, lacking detailed annotations on attributes (e.g., pose and expression), so the influences of different attributes on face recognition have been poorly investigated. In this paper, we address the above-mentioned issues in face recognition using synthetic face images, i.e., SynFace. Specifically, we first explore the performance gap between recent state-of-the-art face recognition models trained with synthetic and real face images. We then analyze the underlying causes behind the performance gap, e.g., the poor intra-class variations and the domain gap between synthetic and real face images. Inspired by this, we devise the SynFace with identity mixup (IM) and domain mixup (DM) to mitigate the above performance gap, demonstrating the great potentials of synthetic data for face recognition. Furthermore, with the controllable face synthesis model, we can easily manage different factors of synthetic face generation, including pose, expression, illumination, the number of identities, and samples per identity. Therefore, we also perform a systematically empirical analysis on synthetic face images to provide some insights on how to effectively utilize synthetic data for face recognition.

****

Camera Distortion-Aware 3D Human Pose Estimation in Video With Optimization-Based Meta-Learning

Hanbyel Cho, Yooshin Cho, Jaemyung Yu, Junmo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11169-11178

Existing 3D human pose estimation algorithms trained on distortion-free datasets suffer performance drop when applied to new scenarios with a specific camera distortion. In this paper, we propose a simple yet effective model for 3D human pose estimation in video that can quickly adapt to any distortion environment by utilizing MAML, a representative optimization-based meta-learning algorithm. We consider a sequence of 2D keypoints in a particular distortion as a single task of MAML. However, due to the absence of a large-scale dataset in a distorted environment, we propose an efficient method to generate synthetic distorted data from undistorted 2D keypoints. For the evaluation, we assume two practical testing situations depending on whether a motion capture sensor is available or not. In particular, we propose Inference Stage Optimization using bone-length symmetry and consistency. Extensive evaluation shows that our proposed method successfully adapts to various degrees of distortion in the testing phase and outperforms the existing state-of-the-art approaches. The proposed method is useful in practice because it does not require camera calibration and additional computations in a testing set-up.

****

KiloNeRF: Speeding Up Neural Radiance Fields With Thousands of Tiny MLPs

Christian Reiser, Songyou Peng, Yiyi Liao, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14335-14345

NeRF synthesizes novel views of a scene with unprecedented quality by fitting a neural radiance field to RGB images. However, NeRF requires querying a deep Multi-Layer Perceptron (MLP) millions of times, leading to slow rendering times, even on modern GPUs. In this paper, we demonstrate that real-time rendering is possible by utilizing thousands of tiny MLPs instead of one single large MLP. In our setting, each individual MLP only needs to represent parts of the scene, thus smaller and faster-to-evaluate MLPs can be used. By combining this divide-and-conquer strategy with further optimizations, rendering is accelerated by three orders of magnitude compared to the original NeRF model without incurring high storage costs. Further, using teacher-student distillation for training, we show that this speed-up can be achieved without sacrificing visual quality.

****

Do Image Classifiers Generalize Across Time?
Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, Ludwig Schmidt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9661-9669
Vision models notoriously flicker when applied to videos: they correctly recognize objects in some frames, but fail on perceptually similar, nearby frames. In this work, we systematically analyze the robustness of image classifiers to such temporal perturbations in videos. To do so, we construct two new datasets, ImageNet-Vid-Robust and YTBB-Robust, containing a total of 57,897 images grouped into 3,139 sets of perceptually similar images. Our datasets were derived from ImageNet-Vid and YouTube-BB, respectively, and thoroughly re-annotated by human experts for image similarity. We evaluate a diverse array of classifiers pre-trained on ImageNet and show a median classification accuracy drop of 16 and 10 points, respectively, on our two datasets. Additionally, we evaluate three detection models and show that natural perturbations induce both classification as well as localization errors, leading to a median drop in detection mAP of 14 points. Our analysis demonstrates that perturbations occurring naturally in videos pose a substantial and realistic challenge to deploying convolutional neural networks in environments that require both reliable and low-latency predictions.
****
Refining Action Segmentation With Hierarchical Video Representations
Hyemin Ahn, Dongheui Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16302-16310
In this paper, we propose Hierarchical Action Segmentation Refiner (HASR), which can refine temporal action segmentation results from various models by understanding the overall context of a given video in a hierarchical way. When a backbone model for action segmentation estimates how the given video can be segmented, our model extracts segment-level representations based on frame-level features, and extracts a video-level representation based on the segment-level representations. Based on these hierarchical representations, our model can refer to the overall context of the entire video, and predict how the segment labels that are out of context should be corrected. Our HASR can be plugged into various action segmentation models (MS-TCN, SSTDA, ASRF), and improve the performance of state-of-the-art models based on three challenging datasets (GTEA, 50Salads, and Breakfast). For example, in 50Salads dataset, the segmental edit score improves from 67.9% to 77.4% (MS-TCN), from 75.8% to 77.3% (SSTDA), from 79.3% to 81.0% (ASRF). In addition, our model can refine the segmentation result from the unseen backbone model, which was not referred to when training HASR. This generalization performance would make HASR be an effective tool for boosting up the existing approaches for temporal action segmentation. Our code is available at https://github.com/cotton-ahn/HASR_iccv2021.
****
Hierarchical Disentangled Representation Learning for Outdoor Illumination Estimation and Editing
Piaopiao Yu, Jie Guo, Fan Huang, Cheng Zhou, Hongwei Che, Xiao Ling, Yanwen Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15313-15322
Data-driven sky models have gained much attention in outdoor illumination prediction recently, showing superior performance against analytical models. However, naively compressing an outdoor panorama into a low-dimensional latent vector, as existing models have done, causes two major problems. One is the mutual interference between the HDR intensity of the sun and the complex textures of the surrounding sky, and the other is the lack of fine-grained control over independent lighting factors due to the entangled representation. To address these issues, we propose a hierarchical disentangled sky model (HDSky) for outdoor illumination prediction. With this model, any outdoor panorama can be hierarchically disentangled into several factors based on three well-designed autoencoders. The first autoencoder compresses each sunny panorama into a sky vector and a sun vector with some constraints. The second autoencoder and the third autoencoder further disentangle the sun intensity and the sky intensity from the sun vector and the sky

vector with several customized loss functions respectively. Moreover, a unified framework is designed to predict all-weather sky information from a single outdoor image. Through extensive experiments, we demonstrate that the proposed model significantly improves the accuracy of outdoor illumination prediction. It also allows users to intuitively edit the predicted panorama (e.g., changing the position of the sun while preserving others), without sacrificing physical plausibility.

****

InSeGAN: A Generative Approach to Segmenting Identical Instances in Depth Images
Anoop Cherian, Gonçalo Dias Pais, Siddarth Jain, Tim K. Marks, Alan Sullivan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10023-10032

In this paper, we present InSeGAN an unsupervised 3D generative adversarial network (GAN) for segmenting (nearly) identical instances of rigid objects in depth images. Using an analysis-by-synthesis approach, we design a novel GAN architecture to synthesize a multiple-instance depth image with independent control over each instance. InSeGAN takes in a set of code vectors (e.g., random noise vectors), each encoding the 3D pose of an object that is represented by a learned implicit object template. The generator has two distinct modules. The first module, the instance feature generator, uses each encoded pose to transform the implicit template into a feature map representation of each object instance. The second module, the depth image renderer, aggregates all of the single-instance feature maps output by the first module and generates a multiple-instance depth image. A discriminator distinguishes the generated multiple-instance depth images from the distribution of true depth images. To use our model for instance segmentation, we propose an instance pose encoder that learns to take in a generated depth image and reproduce the pose code vectors for all of the object instances. To evaluate our approach, we introduce a new synthetic dataset, "Insta-10," consisting of 100,000 depth images each with 5 instances of an object from one of 10 classes. Our experiments on Insta-10, as well as on real-world noisy depth images, show that InSeGAN achieves state-of-the-art performance, often outperforming prior methods by large margins.

****

High-Performance Discriminative Tracking With Transformers
Bin Yu, Ming Tang, Linyu Zheng, Guibo Zhu, Jinqiao Wang, Hao Feng, Xuetao Feng, Hanqing Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9856-9865

End-to-end discriminative trackers improve the state of the art significantly, yet the improvement in robustness and efficiency is restricted by the conventional discriminative model, i.e., least-squares based regression. In this paper, we present DTT, a novel single-object discriminative tracker, based on an encoder-decoder Transformer architecture. By self- and encoder-decoder attention mechanisms, our approach is able to exploit the rich scene information in an end-to-end manner, effectively removing the need for hand-designed discriminative models. In online tracking, given a new test frame, dense prediction is performed at all spatial positions. Not only location, but also bounding box of the target object is obtained in a robust fashion, streamlining the discriminative tracking pipeline. DTT is conceptually simple and easy to implement. It yields state-of-the-art performance on four popular benchmarks including GOT-10k, LaSOT, NfS, and TrackingNet while running at over 50 FPS, confirming its effectiveness and efficiency. We hope DTT may provide a new perspective for single-object visual tracking.

****

GraphFPN: Graph Feature Pyramid Network for Object Detection
Gangming Zhao, Weifeng Ge, Yizhou Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2763-2772

Feature pyramids have been proven powerful in image understanding tasks that require multi-scale features. Stateof-the-art methods for multi-scale feature learning focus on performing feature interactions across space and scales using neural networks with a fixed topology. In this paper, we propose graph feature pyramid networks that are capable of adapting their topological structures to varying

intrinsic image structures, and supporting simultaneous feature interactions acr oss all scales. We first define an image specific superpixel hierarchy for each input image to represent its intrinsic image structures. The graph feature pyram id network inherits its structure from this superpixel hierarchy. Contextual and hierarchical layers are designed to achieve feature interactions within the sam e scale and across different scales, respectively. To make these layers more pow erful, we introduce two types of local channel attention for graph neural networ ks by generalizing global channel attention for convolutional neural networks. T he proposed graph feature pyramid network can enhance the multiscale features fr om a convolutional feature pyramid network. We evaluate our graph feature pyrami d network in the object detection task by integrating it into the Faster RCNN al gorithm. The modified algorithm not only outperforms previous state-of-the-art f eature pyramid based methods with a clear margin but also outperforms other popu lar detection methods on both MS-COCO 2017 validation and test datasets.

****

Self-Supervised 3D Hand Pose Estimation From Monocular RGB via Contrastive Learn ing

Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11 230-11239

Encouraged by the success of contrastive learning on image classification tasks, we propose a new self-supervised method for the structured regression task of 3 D hand pose estimation. Contrastive learning makes use of unlabeled data for the purpose of representation learning via a loss formulation that encourages the l earned feature representations to be invariant under any image transformation. F or 3D hand pose estimation, it too is desirable to have invariance to appearance transformation such as color jitter. However, the task requires equivariance un der affine transformations, such as rotation and translation. To address this is sue, we propose an equivariant contrastive objective and demonstrate its effecti veness in the context of 3D hand pose estimation. We experimentally investigate the impact of invariant and equivariant contrastive objectives and show that lea rning equivariant features leads to better representations for the task of 3D ha nd pose estimation. Furthermore, we show that standard ResNets with sufficient d epth, trained on additional unlabeled data, attain improvements of up to 14.5% i n PA-EPE on FreiHAND and thus achieves state-of-the-art performance without any task specific, specialized architectures. Code and models are available at https ://ait.ethz.ch/projects/2021/PeCLR

****

NeuSpike-Net: High Speed Video Reconstruction via Bio-Inspired Neuromorphic Came ras

Lin Zhu, Jianing Li, Xiao Wang, Tiejun Huang, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2400-2409

Neuromorphic vision sensor is a new bio-inspired imaging paradigm that emerged i n recent years, which continuously sensing luminance intensity and firing asynch ronous spikes (events) with high temporal resolution. Typically, there are two t ypes of neuromorphic vision sensors, namely dynamic vision sensor (DVS) and spik e camera. From the perspective of bio-inspired sampling, DVS only perceives move ment by imitating the retinal periphery, while the spike camera was developed to perceive fine textures by simulating the fovea. It is meaningful to explore how to combine two types of neuromorphic cameras to reconstruct high quality image like human vision. In this paper, we propose a NeuSpike-Net to learn both the hi gh dynamic range and high motion sensitivity of DVS and the full texture samplin g of spike camera to achieve high-speed and high dynamic image reconstruction. W e propose a novel representation to effectively extract the temporal information of spike and event data. By introducing the feature fusion module, the two type s of neuromorphic data achieve complementary to each other. The experimental res ults on the simulated and real datasets demonstrate that the proposed approach i s effective to reconstruct high-speed and high dynamic range images via the comb ination of spike and event data.

****

Admix: Enhancing the Transferability of Adversarial Attacks
Xiaosen Wang, Xuanran He, Jingdong Wang, Kun He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16158-16167
Deep neural networks are known to be extremely vulnerable to adversarial examples under white-box setting. Moreover, the malicious adversaries crafted on the surrogate (source) model often exhibit black-box transferability on other models with the same learning task but having different architectures. Recently, various methods are proposed to boost the adversarial transferability, among which the input transformation is one of the most effective approaches. We investigate in this direction and observe that existing transformations are all applied on a single image, which might limit the adversarial transferability. To this end, we propose a new input transformation based attack method called Admix that considers the input image and a set of images randomly sampled from other categories. Instead of directly calculating the gradient on the original input, Admix calculates the gradient on the input image admixed with a small portion of each add-in image while using the original label of the input to craft more transferable adversaries.
****

ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning
Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, Yale Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10274-10284
The natural association between visual observations and their corresponding sound provides powerful self-supervisory signals for learning video representations, which makes the ever-growing amount of online videos an attractive source of training data. However, large portions of online videos contain irrelevant audio-visual signals because of edited/overdubbed audio, and models trained on such uncurated videos have shown to learn suboptimal representations. Therefore, existing self-supervised approaches rely on datasets with predetermined taxonomies of semantic concepts, where there is a high chance of audio-visual correspondence. Unfortunately, constructing such datasets require labor intensive manual annotation and/or verification, which severely limits the utility of online videos for large-scale learning. In this work, we present an automatic dataset curation approach based on subset optimization where the objective is to maximize the mutual information between audio and visual channels in videos. We demonstrate that our approach finds videos with high audio-visual correspondence and show that self-supervised models trained on our data achieve competitive performances compared to models trained on existing manually curated datasets. The most significant benefit of our approach is scalability: We release ACAV100M that contains 100 million videos with high audio-visual correspondence, ideal for self-supervised video representation learning.
****

Local Temperature Scaling for Probability Calibration
Zhipeng Ding, Xu Han, Peirong Liu, Marc Niethammer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6889-6899
For semantic segmentation, label probabilities are often uncalibrated as they are typically only the by-product of a segmentation task. Intersection over Union (IoU) and Dice score are often used as criteria for segmentation success, while metrics related to label probabilities are not often explored. However, probability calibration approaches have been studied, which match probability outputs with experimentally observed errors. These approaches mainly focus on classification tasks, but not on semantic segmentation. Thus, we propose a learning-based calibration method that focuses on multi-label semantic segmentation. Specifically, we adopt a convolutional neural network to predict local temperature values for probability calibration. One advantage of our approach is that it does not change prediction accuracy, hence allowing for calibration as a post-processing step. Experiments on the COCO, CamVid, and LPBA40 datasets demonstrate improved calibration performance for a range of different metrics. We also demonstrate the good performance of our method for multi-atlas brain segmentation from magnetic r

esonance images.
****
RPVNet: A Deep and Efficient Range-Point-Voxel Fusion Network for LiDAR Point Cloud Segmentation
Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16024-16033
Point clouds can be represented in many forms (views), typically, point-based sets, voxel-based cells or range-based images(i.e., panoramic view). The point-based view is geometrically accurate, but it is disordered, which makes it difficult to find local neighbors efficiently. The voxel-based view is regular, but sparse, and computation grows cubicly when voxel resolution increases. The range-based view is regular and generally dense, however spherical projection makes physical dimensions distorted. Both voxel- and range-based views suffer from quantization loss, especially for voxels when facing large-scale scenes. In order to utilize different view's advantages and alleviate their own shortcomings in fine-grained segmentation task, we propose a novel range-point-voxel fusion network, namely RPVNet. In this network, we devise a deep fusion framework with multiple and mutual information interactions among these three views, and propose a gated fusion module (termed as GFM), which can adaptively merge the three features based on concurrent inputs. Moreover, the proposed RPV interaction mechanism is highly efficient, and we summarize it to a more general formulation. By leveraging this efficient interaction and relatively lower voxel resolution, our method is also proved to be more efficient. Finally, we evaluated the proposed model on two large-scale datasets, i.e., SemanticKITTI and nuScenes, and it shows state-of-the-art performance on both of them. Note that, our method currently ranks 1st on SemanticKITTI leaderboard without any extra tricks.
****
WarpedGANSpace: Finding Non-Linear RBF Paths in GAN Latent Space
Christos Tzelepis, Georgios Tzimiropoulos, Ioannis Patras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6393-6402
This work addresses the problem of discovering, in an unsupervised manner, interpretable paths in the latent space of pretrained GANs, so as to provide an intuitive and easy way of controlling the underlying generative factors. In doing so, it addresses some of the limitations of the state-of-the-art works, namely, a) that they discover directions that are independent of the latent code, i.e., paths that are linear, and b) that their evaluation relies either on visual inspection or on laborious human labeling. More specifically, we propose to learn non-linear warpings on the latent space, each one parametrized by a set of RBF-based latent space warping functions, and where each warping gives rise to a family of non-linear paths via the gradient of the function. Building on the work of Voynov and Babenko, that discovers linear paths, we optimize the trainable parameters of the set of RBFs, so as that images that are generated by codes along different paths, are easily distinguishable by a discriminator network. This leads to easily distinguishable image transformations, such as pose and facial expressions in facial images. We show that linear paths can be derived as a special case of our method, and show experimentally that non-linear paths in the latent space lead to steeper, more disentangled and interpretable changes in the image space than in state-of-the art methods, both qualitatively and quantitatively. We make the code and the pretrained models publicly available at: https://github.com/chi0tzp/WarpedGANSpace.
****
CodeNeRF: Disentangled Neural Radiance Fields for Object Categories
Wonbong Jang, Lourdes Agapito; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12949-12958
CodeNeRF is an implicit 3D neural representation that learns the variation of object shapes and textures across a category and can be trained, from a set of posed images, to synthesize novel views of unseen objects. Unlike the original NeRF, which is scene specific, CodeNeRF learns to disentangle shape and texture by learning separate embeddings. At test time, given a single unposed image of an un

seen object, CodeNeRF jointly estimates camera viewpoint, and shape and appearance codes via optimization. Unseen objects can be reconstructed from a single image, and then rendered from new viewpoints or their shape and texture edited by varying the latent codes. We conduct experiments on the SRN benchmark, which show that CodeNeRF generalises well to unseen objects and achieves on-par performance with methods that require known camera pose at test time. Our results on real-world images demonstrate that CodeNeRF can bridge the sim-to-real gap. Project page: https://github.com/wayne1123/code-nerf

****

Infinite Nature: Perpetual View Generation of Natural Scenes From a Single Image

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, Angjoo Kanazawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14458-14467

We introduce the problem of perpetual view generation - long-range generation of novel views corresponding to an arbitrarily long camera trajectory given a single image. This is a challenging problem that goes far beyond the capabilities of current view synthesis methods, which quickly degenerate when presented with large camera motions. Methods for video generation also have limited ability to produce long sequences and are often agnostic to scene geometry. We take a hybrid approach that integrates both geometry and image synthesis in an iterative render, refine, and repeat framework, allowing for long-range generation that cover large distances after hundreds of frames. Our approach can be trained from a set of monocular video sequences. We propose a dataset of aerial footage of coastal scenes, and compare our method with recent view synthesis and conditional video generation baselines, showing that it can generate plausible scenes for much longer time horizons over large camera trajectories compared to existing methods. Project page at https://infinite-nature.github.io/.

****

Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers

Hila Chefer, Shir Gur, Lior Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 397-406

Transformers are increasingly dominating multi-modal reasoning tasks, such as visual question answering, achieving state-of-the-art results thanks to their ability to contextualize information using the self-attention and co-attention mechanisms. These attention modules also play a role in other computer vision tasks including object detection and image segmentation. Unlike Transformers that only use self-attention, Transformers with co-attention require to consider multiple attention maps in parallel in order to highlight the information that is relevant to the prediction in the model's input. In this work, we propose the first method to explain prediction by any Transformer-based architecture, including bi-modal Transformers and Transformers with co-attentions. We provide generic solutions and apply these to the three most commonly used of these architectures: (i) pure self-attention, (ii) self-attention combined with co-attention, and (iii) encoder-decoder attention. We show that our method is superior to all existing methods which are adapted from single modality explainability.

****

Real-Time Image Enhancer via Learnable Spatial-Aware 3D Lookup Tables

Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, Youliang Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2471-2480

Recently, deep learning-based image enhancement algorithms achieved state-of-the-art (SOTA) performance on several publicly available datasets. However, most existing methods fail to meet practical requirements either for visual perception or for computation efficiency, especially for high-resolution images. In this paper, we propose a novel real-time image enhancer via learnable spatial-aware 3-dimentional lookup tables(3D LUTs), which well considers global scenario and local spatial information. Specifically, we introduce a light weight two-head weight predictor that has two outputs. One is a 1D weight vector used for image-level scenario adaptation, the other is a 3D weight map aimed for pixel-wise category

fusion. We learn the spatial-aware 3D LUTs and fuse them according to the aforem
entioned weights in an end-to-end manner. The fused LUT is then used to transfor
m the source image into the target tone in an efficient way. Extensive results s
how that our model outperforms SOTA image enhancement methods on public datasets
 both subjectively and objectively, and that our model only takes about 4ms to p
rocess a 4K resolution image on one NVIDIA V100 GPU.
****

STAR: A Structure-Aware Lightweight Transformer for Real-Time Image Enhancement
Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, Jinwei Gu; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
1, pp. 4106-4115
Image and video enhancement such as color constancy, low light enhancement, and
tone mapping on smartphones is challenging because high-quality images should be
 achieved efficiently with a limited resource budget. Unlike prior works that ei
ther used very deep CNNs or large Transformer models, we propose a \underline s
eman\underline t ic-\underline a wa\underline r e lightweight Transformer, terme
d STAR, for real-time image enhancement. STAR is formulated to capture long-rang
e dependencies between image patches, which naturally and implicitly captures th
e semantic relationships of different regions in an image. STAR is a general arc
hitecture that can be easily adapted to different image enhancement tasks. Exten
sive experiments show that STAR can effectively boost the quality and efficiency
 of many tasks such as illumination enhancement, auto white balance, and photo r
etouching, which are indispensable components for image processing on smartphone
s. For example, STAR reduces model complexity and improves image quality compare
d to the recent state-of-the-art [??] on the MIT-Adobe FiveK dataset [??] (i.e.,
 1.8dB PSNR improvements with 25% parameters and 13% float operations.)
****

Continuous Copy-Paste for One-Stage Multi-Object Tracking and Segmentation
Zhenbo Xu, Ajin Meng, Zhenbo Shi, Wei Yang, Zhi Chen, Liusheng Huang; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
15323-15332
Current one-step multi-object tracking and segmentation (MOTS) methods lag behin
d recent two-step methods. By separating the instance segmentation stage from th
e tracking stage, two-step methods can exploit non-video datasets as extra data
for training instance segmentation. Moreover, instances belonging to different I
Ds on different frames, rather than limited numbers of instances in raw consecut
ive frames, can be gathered to allow more effective hard example mining in the t
raining of trackers. In this paper, we bridge this gap by presenting a novel dat
a augmentation strategy named continuous copy-paste (CCP). Our intuition behind
CCP is to fully exploit the pixel-wise annotations provided by MOTS to actively
increase the number of instances as well as unique instance IDs in training. Wit
hout any modifications to frameworks, current MOTS methods achieve significant p
erformance gains when trained with CCP. Based on CCP, we propose the first effec
tive one-stage online MOTS method named CCPNet, which generates instance masks a
s well as the tracking results in one shot. Our CCPNet surpasses all state-of-th
e-art methods by large margins (3.8% higher sMOTSA and 4.1% higher MOTSA for ped
estrians on the KITTI MOTS Validation) and ranks 1st on the KITTI MOTS leaderboa
rd. Evaluations across three datasets also demonstrate the effectiveness of both
 CCP and CCPNet. Our codes are publicly available at: https://github.com/detectR
ecog/CCP.
****

Hand-Object Contact Consistency Reasoning for Human Grasps Generation
Hanwen Jiang, Shaowei Liu, Jiashun Wang, Xiaolong Wang; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2021, pp. 11107-11116
While predicting robot grasps with parallel jaw grippers have been well studied
and widely applied in robot manipulation tasks, the study on natural human grasp
 generation with a multi-finger hand remains a very challenging problem. In this
 paper, we propose to generate human grasps given a 3D object in the world. Our
key observation is that it is crucial to model the consistency between the hand
contact points and object contact regions. That is, we encourage the prior hand

contact points to be close to the object surface and the object common contact r
egions to be touched by the hand at the same time. Based on the hand-object cont
act consistency, we design novel objectives in training the human grasp generati
on model and also a new self-supervised task which allows the grasp generation n
etwork to be adjusted even during test time. Our experiments show significant im
provement in human grasp generation over state-of-the-art approaches by a large
margin. More interestingly, by optimizing the model during test time with the se
lf-supervised task, it helps achieve larger gain on unseen and out-of-domain obj
ects. Project page: https://hwjiang1510.github.io/GraspTTA/.
****

FashionMirror: Co-Attention Feature-Remapping Virtual Try-On With Sequential Tem
plate Poses
Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, Wen-Huang Cheng; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 13809-13818
Virtual try-on tasks have drawn increased attention. Prior arts focus on tacklin
g this task via warping clothes and fusing the information at the pixel level wi
th the help of semantic segmentation. However, conducting semantic segmentation
is time-consuming and easily causes error accumulation over time. Besides, warpi
ng the information at the pixel level instead of the feature level limits the pe
rformance (e.g., unable to generate different views) and is unstable since it di
rectly demonstrates the results even with a misalignment. In contrast, fusing in
formation at the feature level can be further refined by the convolution to obta
in the final results. Based on these assumptions, we propose a co-attention feat
ure-remapping framework, namely FashionMirror, that generates the try-on results
according to the driven-pose sequence in two stages. In the first stage, we con
sider the source human image and the target try-on clothes to predict the remove
d mask and the try-on clothing mask, which replaces the pre-processed semantic s
egmentation and reduces the inference time. In the second stage, we first remove
the clothes on the source human via the removed mask and warp the clothing feat
ures conditioning on the try-on clothing mask to fit the next frame human. Meanw
hile, we predict the optical flows from the consecutive 2D poses and warp the so
urce human to the next frame at the feature level. Then, we enhance the clothing
features and source human features in every frame to generate realistic try-on
results with spatio-temporal smoothness. Both qualitative and quantitative resul
ts show that FashionMirror outperforms the state-of-the-art virtual try-on appro
aches.
****

Reconcile Prediction Consistency for Balanced Object Detection
Keyang Wang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2021, pp. 3631-3640
Classification and regression are two pillars of object detectors. In most CNN-b
ased detectors, these two pillars are optimized independently. Without direct in
teractions between them, the classification loss and the regression loss can not
be optimized synchronously toward the optimal direction in the training phase.
This clearly leads to lots of inconsistent predictions with high classification
score but low localization accuracy or low classification score but high localiz
ation accuracy in the inference phase, especially for the objects of irregular s
hape and occlusion, which severely hurts the detection performance of existing d
etectors after NMS. To reconcile prediction consistency for balanced object dete
ction, we propose a Harmonic loss to harmonize the optimization of classificatio
n branch and localization branch. The Harmonic loss enables these two branches t
o supervise and promote each other during training, thereby producing consistent
predictions with high co-occurrence of top classification and localization in t
he inference phase. Furthermore, in order to prevent the localization loss from
being dominated by outliers during training phase, a Harmonic IoU loss is propos
ed to harmonize the weight of the localization loss of different IoU-level sampl
es. Comprehensive experiments on benchmarks PASCAL VOC and MS COCO demonstrate t
he generality and effectiveness of our model for facilitating existing object de
tectors to state-of-the-art accuracy.

****
Confidence Calibration for Domain Generalization Under Covariate Shift
Yunye Gong, Xiao Lin, Yi Yao, Thomas G. Dietterich, Ajay Divakaran, Melinda Gervasio; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8958-8967
Existing calibration algorithms address the problem of covariate shift via unsupervised domain adaptation. However, these methods suffer from the following limitations: 1) they require unlabeled data from the target domain, which may not be available at the stage of calibration in real-world applications and 2) their performance depends heavily on the disparity between the distributions of the source and target domains. To address these two limitations, we present novel calibration solutions via domain generalization. Our core idea is to leverage multiple calibration domains to reduce the effective distribution disparity between the target and calibration domains for improved calibration transfer without needing any data from the target domain. We provide theoretical justification and empirical experimental results to demonstrate the effectiveness of our proposed algorithms. Compared against state-of-the-art calibration methods designed for domain adaptation, we observe a decrease of 8.86 percentage points in expected calibration error or, equivalently, an increase of 35 percentage points in improvement ratio for multi-class classification on the Office-Home dataset.
****
Self-Supervised Video Representation Learning With Meta-Contrastive Network
Yuanze Lin, Xun Guo, Yan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8239-8249
Self-supervised learning has been successfully applied to pre-train video representations, which aims at efficient adaptation from pre-training domain to downstream tasks. Existing approaches merely leverage contrastive loss to learn instance-level discrimination. However, lack of category information will lead to hard-positive problem that constrains the generalization ability of this kind of methods. We find that the multi-task process of meta learning can provide a solution to this problem. In this paper, we propose a Meta-Contrastive Network (MCN), which combines the contrastive learning and meta learning, to enhance the learning ability of existing self-supervised approaches. Our method contains two training stages based on model-agnostic meta learning (MAML), each of which consists of a contrastive branch and a meta branch. Extensive evaluations demonstrate the effectiveness of our method. For two downstream tasks, i.e., video action recognition and video retrieval, MCN outperforms state-of-the-art approaches on UCF101 and HMDB51 datasets. To be more specific, with R(2+1)D backbone, MCN achieves Top-1 accuracies of 84.8% and 54.5% for video action recognition, as well as 52.5% and 23.7% for video retrieval.
****
A Confidence-Based Iterative Solver of Depths and Surface Normals for Deep Multi-View Stereo
Wang Zhao, Shaohui Liu, Yi Wei, Hengkai Guo, Yong-Jin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6168-6177
In this paper, we introduce a deep multi-view stereo (MVS) system that jointly predicts depths, surface normals and per-view confidence maps. The key to our approach is a novel solver that iteratively solves for per-view depth map and normal map by optimizing an energy potential based upon the local planar assumption. Specifically, the algorithm updates depth map by propagating from neighboring pixels with slanted planes, and updates normal map with local probabilistic plane fitting. Both two steps are monitored by a customized confidence map. This confidence-based solver is not only effective as a post-processing tool for plane based depth refinement and completion, but also differentiable such that it can be efficiently integrated into deep learning pipelines. Our multi-view stereo system employs multiple optimization steps of the solver over the initial prediction of depths and surface normals. The whole system can be trained end-to-end, decoupling the challenging problem of matching pixels within poorly textured regions from the cost volume based neural network. Experimental results on ScanNet and RGB-D Scenes V2 demonstrate state-of-the-art performance of the proposed deep MVS

system on multi-view depth estimation, with our proposed solver consistently improving the depth quality over both conventional and deep learning based MVS pipelines.

****

Unsupervised Depth Completion With Calibrated Backprojection Layers
Alex Wong, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12747-12756

We propose a deep neural network architecture to infer dense depth from an image and a sparse point cloud. It is trained using a video stream and corresponding synchronized sparse point cloud, as obtained from a LIDAR or other range sensor, along with the intrinsic calibration parameters of the camera. At inference time, the calibration of the camera, which can be different than the one used for training, is fed as an input to the network along with the sparse point cloud and a single image. A Calibrated Backprojection Layer backprojects each pixel in the image to three-dimensional space using the calibration matrix and a depth feature descriptor. The resulting 3D positional encoding is concatenated with the image descriptor and the previous layer output to yield the input to the next layer of the encoder. A decoder, exploiting skip-connections, produces a dense depth map. The resulting Calibrated Backprojection Network, or KBNet, is trained without supervision by minimizing the photometric reprojection error. KBNet imputes missing depth value based on the training set, rather than on generic regularization. We test KBNet on public depth completion benchmarks, where it outperforms the state of the art by 30% indoor and 8% outdoor when the same camera is used for training and testing. When the test camera is different, the improvement reaches 62%.

****

Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval
Max Bain, Arsha Nagrani, Gül Varol, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1728-1738

Our objective in this work is video-text retrieval - in particular a joint embedding that enables efficient text-to-video retrieval. The challenges in this area include the design of the visual architecture and the nature of the training data, in that the available large scale video-text training datasets, such as HowTo100M, are noisy and hence competitive performance is achieved only at scale through large amounts of compute. We address both these challenges in this paper. We propose an end-to-end trainable model that is designed to take advantage of both large-scale image and video captioning datasets. Our model is an adaptation and extension of the recent ViT and Timesformer architectures, and consists of attention in both space and time. The model is flexible and can be trained on both image and video text datasets, either independently or in conjunction. It is trained with a curriculum learning schedule that begins by treating images as 'frozen' snapshots of video, and then gradually learns to attend to increasing temporal context when trained on video datasets. We also provide a new video-text pretraining dataset WebVid-2M, comprised of over two million videos with weak captions scraped from the internet. Despite training on datasets that are an order of magnitude smaller, we show that this approach yields state-of-the-art results on standard downstream video-retrieval benchmarks including MSR-VTT, DiDeMo and MSVD.

****

LIRA: Learnable, Imperceptible and Robust Backdoor Attacks
Khoa Doan, Yingjie Lao, Weijie Zhao, Ping Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11966-11976

Recently, machine learning models have demonstrated to be vulnerable to backdoor attacks, primarily due to the lack of transparency in black-box models such as deep neural networks. A third-party model can be poisoned such that it works adequately in normal conditions but behaves maliciously on samples with specific trigger patterns. However, the trigger injection function is manually defined in most existing backdoor attack methods, e.g., placing a small patch of pixels on an image or slightly deforming the image before poisoning the model. This results in a two-stage approach with a sub-optimal attack success rate and a lack of co

mplete stealthiness under human inspection. In this paper, we propose a novel an d stealthy backdoor attack framework, LIRA, which jointly learns the optimal, st ealthy trigger injection function and poisons the model. We formulate such an ob jective as a non-convex, constrained optimization problem. Under this optimizati on framework, the trigger generator function will learn to manipulate the input with imperceptible noise to preserve the model performance on the clean data and maximize the attack success rate on the poisoned data. Then, we solve this chal lenging optimization problem with an efficient, two-stage stochastic optimizatio n procedure. Finally, the proposed attack framework achieves 100% success rates in several benchmark datasets, including MNIST, CIFAR10, GTSRB, and T-ImageNet, while simultaneously bypassing existing backdoor defense methods and human inspe ction.

****

DnD: Dense Depth Estimation in Crowded Dynamic Indoor Scenes

Dongki Jung, Jaehoon Choi, Yonghan Lee, Deokhwa Kim, Changick Kim, Dinesh Manoch a, Donghwan Lee; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2021, pp. 12797-12807

We present a novel approach for estimating depth from a monocular camera as it m oves through complex and crowded indoor environments, e.g., a department store o r a metro station. Our approach predicts absolute scale depth maps over the enti re scene consisting of a static background and multiple moving people, by traini ng on dynamic scenes. Since it is difficult to collect dense depth maps from cro wded indoor environments, we design our training framework without requiring gro undtruth depths produced from depth sensing devices. Our network leverages RGB i mages and sparse depth maps generated from traditional 3D reconstruction methods to estimate dense depth maps. We use two constraints to handle depth for non-ri gidly moving people without tracking their motion explicitly. We demonstrate tha t our approach offers consistent improvements over recent depth estimation metho ds on the NAVERLABS dataset, which includes complex and crowded scenes.

****

Click To Move: Controlling Video Generation With Sparse Motion

Pierfrancesco Ardino, Marco De Nadai, Bruno Lepri, Elisa Ricci, Stéphane Lathuil ière; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2021, pp. 14749-14758

This paper introduces Click to Move (C2M), a novel framework for video generatio n where the user can control the motion of the synthesized video through mouse c licks specifying simple object trajectories of the key objects in the scene. Our model receives as input an initial frame, its corresponding segmentation map an d the sparse motion vectors encoding the input provided by the user. It outputs a plausible video sequence starting from the given frame and with a motion that is consistent with user input. Notably, our proposed deep architecture incorpora tes a Graph Convolution Network (GCN) modelling the movements of all the objects in the scene in a holistic manner and effectively combining the sparse user mot ion information and image features. Experimental results show that C2M outperfor ms existing methods on two publicly available datasets, thus demonstrating the e ffectiveness of our GCN framework at modelling object interactions. The source c ode is publicly available at https://github.com/PierfrancescoArdino/C2M.

****

Towards Mixed-Precision Quantization of Neural Networks via Constrained Optimiza tion

Weihan Chen, Peisong Wang, Jian Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5350-5359

Quantization is a widely used technique to compress and accelerate deep neural n etworks. However, conventional quantization methods use the same bit-width for a ll (or most of) the layers, which often suffer significant accuracy degradation in the ultra-low precision regime and ignore the fact that emergent hardware acc elerators begin to support mixed-precision computation. Consequently, we present a novel and principled framework to solve the mixed-precision quantization prob lem in this paper. Briefly speaking, we first formulate the mixed-precision quan tization as a discrete constrained optimization problem. Then, to make the optim

ization tractable, we approximate the objective function with second-order Taylo
r expansion and propose an efficient approach to compute its Hessian matrix. Fin
ally, based on the above simplification, we show that the original problem can b
e reformulated as a Multiple Choice Knapsack Problem (MCKP) and propose a greedy
 search algorithm to solve it efficiently. Compared with existing mixed-precisio
n quantization works, our method is derived in a principled way and much more co
mputationally efficient. Moreover, extensive experiments conducted on the ImageN
et dataset and various kinds of network architectures also demonstrate its super
iority over existing uniform and mixed-precision quantization approaches.
****

Dual-Camera Super-Resolution With Aligned Attention Modules
Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, Qifeng Chen; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2001-201
0
We present a novel approach to reference-based super-resolution (RefSR) with the
 focus on dual-camera super-resolution (DCSR), which utilizes reference images f
or high-quality and high-fidelity results. Our proposed method generalizes the s
tandard patch-based feature matching with spatial alignment operations. We furth
er explore the dual-camera super-resolution that is one promising application of
 RefSR, and build a dataset that consists of 146 image pairs from the main and t
elephoto cameras in a smartphone. To bridge the domain gaps between real-world i
mages and the training images, we propose a self-supervised domain adaptation st
rategy for real-world images. Extensive experiments on our dataset and a public
benchmark demonstrate clear improvement achieved by our method over state of the
 art in both quantitative evaluation and visual comparisons.
****

NASOA: Towards Faster Task-Oriented Online Fine-Tuning With a Zoo of Models
Hang Xu, Ning Kang, Gengwei Zhang, Chuanlong Xie, Xiaodan Liang, Zhenguo Li; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
1, pp. 5097-5106
Fine-tuning from pre-trained ImageNet models has been a simple, effective, and p
opular approach for various computer vision tasks. The common practice of fine-t
uning is to adopt a default hyperparameter setting with a fixed pre-trained mode
l, while both of them are not optimized for specific tasks and time constraints.
 Moreover, in cloud computing or GPU clusters where the tasks arrive sequentiall
y in a stream, faster online fine-tuning is a more desired and realistic strateg
y for saving money, energy consumption, and CO2 emission. In this paper, we prop
ose a joint Neural Architecture Search and Online Adaption framework named NASOA
 towards a faster task-oriented fine-tuning upon the request of users. Specifica
lly, NASOA first adopts an offline NAS to identify a group of training-efficient
 networks to form a pretrained model zoo. We propose a novel joint block and mac
ro level search space to enable a flexible and efficient search. Then, by estima
ting fine-tuning performance via an adaptive model by accumulating experience fr
om the past tasks, an online schedule generator is proposed to pick up the most
suitable model and generate a personalized training regime with respect to each
desired task in a one-shot fashion. The resulting model zoo is more training eff
icient than SOTA NAS models, e.g. 6x faster than RegNetY-16GF, and 1.7x faster t
han EfficientNetB3. Experiments on multiple datasets also show that NASOA achiev
es much better fine-tuning results, i.e. improving around 2.1% accuracy than the
 best performance in RegNet series under various time constraints and tasks; 40x
 faster compared to the BOHB method.
****

RandomRooms: Unsupervised Pre-Training From Synthetic Shapes and Randomized Layo
uts for 3D Object Detection
Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, Jie Zhou; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3
283-3292
3D point cloud understanding has made great progress in recent years. However, o
ne major bottleneck is the scarcity of annotated real datasets, especially compa
red to 2D object detection tasks, since a large amount of labor is involved in a

nnotating the real scans of a scene. A promising solution to this problem is to make better use of the synthetic dataset, which consists of CAD object models, to boost the learning on real datasets. This can be achieved by the pre-training and fine-tuning procedure. However, recent work on 3D pre-training exhibits failure when transfer features learned on synthetic objects to other real-world applications. In this work, we put forward a new method called RandomRooms to accomplish this objective. In particular, we propose to generate random layouts of a scene by making use of the objects in the synthetic CAD dataset and learn the 3D scene representation by applying object-level contrastive learning on two random scenes generated from the same set of synthetic objects. The model pre-trained in this way can serve as a better initialization when later fine-tuning on the 3D object detection task. Empirically, we show consistent improvement in downstream 3D detection tasks on several base models, especially when less training data are used, which strongly demonstrates the effectiveness and generalization of our method. Benefiting from the rich semantic knowledge and diverse objects from synthetic data, our method establishes the new state-of-the-art on widely-used 3D detection benchmarks ScanNetV2 and SUN RGB-D. We expect our attempt to provide a new perspective for bridging object and scene-level 3D understanding.
****

From Continuity to Editability: Inverting GANs With Consecutive Images
Yangyang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, Shengfeng He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13910-13918
Existing GAN inversion methods are stuck in a paradox that the inverted codes can either achieve high-fidelity reconstruction, or retain the editing capability. Having only one of them clearly cannot realize real image editing. In this paper, we resolve this paradox by introducing consecutive images (e.g., video frames or the same person with different poses) into the inversion process. The rationale behind our solution is that the continuity of consecutive images leads to inherent editable directions. This inborn property is used for two unique purposes: 1) regularizing the joint inversion process, such that each of the inverted codes is semantically accessible from one of the other and fastened in an editable domain; 2) enforcing inter-image coherence, such that the fidelity of each inverted code can be maximized with the complement of other images. Extensive experiments demonstrate that our alternative significantly outperforms state-of-the-art methods in terms of reconstruction fidelity and editability on both the real image dataset and synthesis dataset. Furthermore, our method provides the first support of video-based GAN inversion and an interesting application of unsupervised semantic transfer from consecutive images. Source code can be found at: https://github.com/cnnlstm/InvertingGANs_with_ConsecutiveImgs.
****

GyroFlow: Gyroscope-Guided Unsupervised Optical Flow Learning
Haipeng Li, Kunming Luo, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12869-12878
Existing optical flow methods are erroneous in challenging scenes, such as fog, rain, and night because the basic optical flow assumptions such as brightness and gradient constancy are broken. To address this problem, we present an unsupervised learning approach that fuses gyroscope into optical flow learning. Specifically, we first convert gyroscope readings into motion fields named gyro field. Second, we design a self-guided fusion module to fuse the background motion extracted from the gyro field with the optical flow and guide the network to focus on motion details. To the best of our knowledge, this is the first deep learning-based framework that fuses gyroscope data and image content for optical flow learning. To validate our method, we propose a new dataset that covers regular and challenging scenes. Experiments show that our method outperforms the state-of-art methods in both regular and challenging scenes. Code and dataset are available at https://github.com/megvii-research/GyroFlow.
****

Towards Discovery and Attribution of Open-World GAN Generated Images
Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, Abhinav Shrivastava; Proceed

ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14094-14103
With the recent progress in Generative Adversarial Networks (GANs), it is imperative for media and visual forensics to develop detectors which can identify and attribute images to the model generating them. Existing works have shown to attribute images to their corresponding GAN sources with high accuracy. However, these works are limited to a closed set scenario, failing to generalize to GANs unseen during train time and are therefore, not scalable with a steady influx of new GANs. We present an iterative algorithm for discovering images generated from previously unseen GANs by exploiting the fact that all GANs leave distinct fingerprints on their generated images. Our algorithm consists of multiple components including network training, out-of-distribution detection, clustering, merge and refine steps. Through extensive experiments, we show that our algorithm discovers unseen GANs with high accuracy and also generalizes to GANs trained on unseen real datasets. We additionally apply our algorithm to attribution and discovery of GANs in an online fashion as well as to the more standard task of real/fake detection. Our experiments demonstrate the effectiveness of our approach to discover new GANs and can be used in an open-world setup.
****

Vector Neurons: A General Framework for SO(3)-Equivariant Networks
Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12200-12209
Invariance and equivariance to the rotation group have been widely discussed in the 3D deep learning community for pointclouds. Yet most proposed methods either use complex mathematical tools that may limit their accessibility, or are tied to specific input data types and network architectures. In this paper, we introduce a general framework built on top of what we call Vector Neuron representations for creating SO(3)-equivariant neural networks for pointcloud processing. Extending neurons from 1D scalars to 3D vectors, our vector neurons enable a simple mapping of SO(3) actions to latent spaces thereby providing a framework for building equivariance in common neural operations -- including linear layers, non-linearities, pooling, and normalizations. Due to their simplicity, vector neurons are versatile and, as we demonstrate, can be incorporated into diverse network architecture backbones, allowing them to process geometry inputs in arbitrary poses. Despite its simplicity, our method performs comparably well in accuracy and generalization with other more complex and specialized state-of-the-art methods on classification and segmentation tasks. We also show for the first time a rotation equivariant reconstruction network. Source code is available at https://github.com/FlyingGiraffe/vnn.
****

Integer-Arithmetic-Only Certified Robustness for Quantized Neural Networks
Haowen Lin, Jian Lou, Li Xiong, Cyrus Shahabi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7828-7837
Adversarial data examples have drawn significant attention from the machine learning and security communities. A line of work on tackling adversarial examples is certified robustness via randomized smoothing that can provide a theoretical robustness guarantee. However, such a mechanism usually uses floating-point arithmetic for calculations in inference and requires large memory footprints and daunting computational costs. These defensive models cannot run efficiently on edge devices nor be deployed on integer-only logical units such as Turing Tensor Cores or integer-only ARM processors. To overcome these challenges, we propose an integer randomized smoothing approach with quantization to convert any classifier into a new smoothed classifier, which uses integer-only arithmetic for certified robustness against adversarial perturbations. We prove a tight robustness guarantee under L2-norm for the proposed approach. We show our approach can obtain a comparable accuracy and 4x 5x speedup over floating-point arithmetic certified robust methods on general-purpose CPUs and mobile devices on two distinct datasets (CIFAR-10 and Caltech-101).
****

Video-Based Person Re-Identification With Spatial and Temporal Memory Networks
Chanho Eom, Geon Lee, Junghyup Lee, Bumsub Ham; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12036-12045
Video-based person re-identification (reID) aims to retrieve person videos with the same identity as a query person across multiple cameras. Spatial and temporal distractors in person videos, such as background clutter and partial occlusions over frames, respectively, make this task much more challenging than image-based person reID. We observe that spatial distractors appear consistently in a particular location, and temporal distractors show several patterns, e.g., partial occlusions occur in the first few frames, where such patterns provide informative cues for predicting which frames to focus on (i.e., temporal attentions). Based on this, we introduce a novel Spatial and Temporal Memory Networks (STMN). The spatial memory stores features for spatial distractors that frequently emerge across video frames, while the temporal memory saves attentions which are optimized for typical temporal patterns in person videos. We leverage the spatial and temporal memories to refine frame-level person representations and to aggregate the refined frame-level features into a sequence-level person representation, respectively, effectively handling spatial and temporal distractors in person videos. We also introduce a memory spread loss preventing our model from addressing particular items only in the memories. Experimental results on standard benchmarks, including MARS, DukeMTMC-VideoReID, and LS-VID, demonstrate the effectiveness of our method.
****

Conformer: Local Features Coupling Global Representations for Visual Recognition
Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, Qixiang Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 367-376
Within Convolutional Neural Network (CNN), the convolution operations are good at extracting local features but experience difficulty to capture global representations. Within visual transformer, the cascaded self-attention modules can capture long-distance feature dependencies but unfortunately deteriorate local feature details. In this paper, we propose a hybrid network structure, termed Conformer, to take advantage of convolutional operations and self-attention mechanisms for enhanced representation learning. Conformer roots in the Feature Coupling Unit (FCU), which fuses local features and global representations under different resolutions in an interactive fashion. Conformer adopts a concurrent structure so that local features and global representations are retained to the maximum extent. Experiments show that Conformer, under the comparable parameter complexity, outperforms the visual transformer (DeiT-B) by 2.3% on ImageNet. On MSCOCO, it outperforms ResNet-101 by 3.7% and 3.6% mAPs for object detection and instance segmentation, respectively, demonstrating the great potential to be a general backbone network. Code is available at github.com/pengzhiliang/Conformer.
****

Lightweight Multi-Person Total Motion Capture Using Sparse Multi-View Cameras
Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, Yebin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5560-5569
Multi-person total motion capture is extremely challenging when it comes to handle severe occlusions, different reconstruction granularities from body to face and hands, drastically changing observation scales and fast body movements. To overcome these challenges above, we contribute a lightweight total motion capture system for multi-person interactive scenarios using only sparse multi-view cameras. By contributing a novel hand and face bootstrapping algorithm, our method is capable of efficient localization and accurate association of the hands and faces even on severe occluded occasions. We leverage both pose regression and keypoints detection methods and further propose a unified two-stage parametric fitting method for achieving pixel-aligned accuracy. Moreover, for extremely self-occluded poses and close interactions, a novel feedback mechanism is proposed to propagate the pixel-aligned reconstructions into the next frame for more accurate association. Overall, we propose the first light-weight total capture system and

achieves fast, robust and accurate multi-person total motion capture performance. The results and experiments show that our method achieves more accurate results than existing methods under sparse-view setups.
****

AGKD-BML: Defense Against Adversarial Attack by Attention Guided Knowledge Distillation and Bi-Directional Metric Learning
Hong Wang, Yuefan Deng, Shinjae Yoo, Haibin Ling, Yuewei Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7658-7667
While deep neural networks have shown impressive performance in many tasks, they are fragile to carefully designed adversarial attacks. We propose a novel adversarial training-based model by Attention Guided Knowledge Distillation and Bi-directional Metric Learning (AGKD-BML). The attention knowledge is obtained from a weight-fixed model trained on a clean dataset, referred to as a teacher model, and transferred to a model that is under training on adversarial examples (AEs), referred to as a student model. In this way, the student model is able to focus on the correct region, as well as correcting the intermediate features corrupted by AEs to eventually improve the model accuracy. Moreover, to efficiently regularize the representation in feature space, we propose a bidirectional metric learning. Specifically, given a clean image, it is first attacked to its most confusing class to get the forward AE. A clean image in the most confusing class is then randomly picked and attacked back to the original class to get the backward AE. A triplet loss is then used to shorten the representation distance between original image and its AE, while enlarge that between the forward and backward AEs. We conduct extensive adversarial robustness experiments on two widely used datasets with different attacks. Our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches. The code of AGKD-BML will be available at: https://github.com/hongw579/AGKD-BML.
****

Revealing the Reciprocal Relations Between Self-Supervised Stereo and Monocular Depth Estimation
Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, Liusheng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15529-15538
Current self-supervised depth estimation algorithms mainly focus on either stereo or monocular only, neglecting the reciprocal relations between them. In this paper, we propose a simple yet effective framework to improve both stereo and monocular depth estimation by leveraging the underlying complementary knowledge of the two tasks. Our approach consists of three stages. In the first stage, the proposed stereo matching network termed StereoNet is trained on image pairs in a self-supervised manner. Second, we introduce an occlusion-aware distillation (OA Distillation) module, which leverages the predicted depths from StereoNet in non-occluded regions to train our monocular depth estimation network named SingleNet. At last, we design an occlusion-aware fusion module (OA Fusion), which generates more reliable depths by fusing estimated depths from StereoNet and SingleNet given the occlusion map. Furthermore, we also take the fused depths as pseudo labels to supervise StereoNet in turn, which brings StereoNet's performance to a new height. Extensive experiments on KITTI dataset demonstrate the effectiveness of our proposed framework. We achieve new SOTA performance on both stereo and monocular depth estimation tasks.
****

MGSampler: An Explainable Sampling Strategy for Video Action Recognition
Yuan Zhi, Zhan Tong, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1513-1522
Frame sampling is a fundamental problem in video action recognition due to the essential redundancy in time and limited computation resources. The existing sampling strategy often employs a fixed frame selection and lacks the flexibility to deal with complex variations in videos. In this paper, we present a simple, sparse, and explainable frame sampler, termed as Motion-Guided Sampler (MGSampler). Our basic motivation is that motion is an important and universal signal that c

an drive us to adaptively select frames from videos. Accordingly, we propose two important properties in our MGSampler design: motion sensitive and motion uniform. First, we present two different motion representations to enable us to efficiently distinguish the motion-salient frames from the background. Then, we devise a motion-uniform sampling strategy based on the cumulative motion distribution to ensure the sampled frames evenly cover all the important segments with high motion salience. Our MGSampler yields a new principled and holistic sample scheme, that could be incorporated into any existing video architecture. Experiments on five benchmarks demonstrate the effectiveness of our MGSampler over previous fixed sampling strategies, and its generalization power across different backbones, video models, and datasets.

****

Robust 2D/3D Vehicle Parsing in Arbitrary Camera Views for CVIS

Hui Miao, Feixiang Lu, Zongdai Liu, Liangjun Zhang, Dinesh Manocha, Bin Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15631-15640

We present a novel approach to robustly detect and perceive vehicles in different camera views as part of a cooperative vehicle-infrastructure system (CVIS). Our formulation is designed for arbitrary camera views and makes no assumptions about intrinsic or extrinsic parameters. First, to deal with multi-view data scarcity, we propose a part-assisted novel view synthesis algorithm for data augmentation. We train a part-based texture inpainting network in a self-supervised manner. Then we render the textured model into the background image with the target 6-DoF pose. Second, to handle various camera parameters, we present a new method that produces dense mappings between image pixels and 3D points to perform robust 2D/3D vehicle parsing. Third, we build the first CVIS dataset for benchmarking, which annotates more than 1540 images (14017 instances) from real-world traffic scenarios. We combine these novel algorithms and datasets to develop a robust approach for 2D/3D vehicle parsing for CVIS. In practice, our approach outperforms SOTA methods on 2D detection, instance segmentation, and 6-DoF pose estimation by 3.8%, 4.3%, and 2.9%, respectively.

****

Recurrent Mask Refinement for Few-Shot Medical Image Segmentation

Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, Xiaohui Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3918-3928

Although having achieved great success in medical image segmentation, deep convolutional neural networks usually require a large dataset with manual annotations for training and are difficult to generalize to unseen classes. Few-shot learning has the potential to address these challenges by learning new classes from only a few labeled examples. In this work, we propose a new framework for few-shot medical image segmentation based on prototypical networks. Our innovation lies in the design of two key modules: 1) a context relation encoder (CRE) that uses correlation to capture local relation features between foreground and background regions; and 2) a recurrent mask refinement module that repeatedly uses the CRE and a prototypical network to recapture the change of context relationship and refine the segmentation mask iteratively. Experiments on two abdomen CT datasets and an abdomen MRI dataset show the proposed method obtains substantial improvement over the state-of-the-art methods by an average of 16.32%, 8.45% and 6.24% in terms of DSC, respectively. Code is publicly available.

****

ECACL: A Holistic Framework for Semi-Supervised Domain Adaptation

Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8578-8587

This paper studies Semi-Supervised Domain Adaptation (SSDA), a practical yet under-investigated research topic that aims to learn a model of good performance using unlabeled samples and a few labeled samples in the target domain, with the help of labeled samples from a source domain. Several SSDA methods have been proposed recently, which however fail to fully exploit the value of the few labeled target samples. In this paper, we propose Enhanced Categorical Alignment and Con

sistency Learning (ECACL), a holistic SSDA framework that incorporates multiple mutually complementary domain alignment techniques. ECACL includes two categoric al domain alignment techniques that achieve class-level alignment, a strong data augmentation based technique that enhances the model's generalizability and a c onsistency learning based technique that forces the model to be robust with imag e perturbations. These techniques are applied on one or multiple of the three in puts (labeled source, unlabeled target, and labeled target) and align the domain s from different perspectives. ECACL unifies them together and achieves fairly c omprehensive domain alignments that are much better than the existing methods: F or example, ECACL raises the state-of-the-art accuracy from 68.4 to 81.1 on VisD A2017 and from 45.5 to 53.4 on DomainNet for the 1-shot setting. Our code is ava ilable at https://github.com/kailigo/pacl.

****

WaveFill: A Wavelet-Based Generation Network for Image Inpainting

Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, Chunyan Miao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14114-14123

Image inpainting aims to complete the missing or corrupted regions of images wit h realistic contents. The prevalent approaches adopt a hybrid objective of recon struction and perceptual quality by using generative adversarial networks. Howev er, the reconstruction loss and adversarial loss focus on synthesizing contents of different frequencies and simply applying them together often leads to inter- frequency conflicts and compromised inpainting. This paper presents WaveFill, a wavelet-based inpainting network that decomposes images into multiple frequency bands and fills the missing regions in each frequency band separately and explic itly. WaveFill decomposes images by using discrete wavelet transform (DWT) that preserves spatial information naturally. It applies L1 reconstruction loss to th e decomposed low-frequency bands and adversarial loss to high-frequency bands, h ence effectively mitigate inter-frequency conflicts while completing images in s patial domain. To address the inpainting inconsistency in different frequency ba nds and fuse features with distinct statistics, we design a novel normalization scheme that aligns and fuses the multi-frequency features effectively. Extensive experiments over multiple datasets show that WaveFill achieves superior image i npainting qualitatively and quantitatively.

****

Egocentric Pose Estimation From Human Vision Span

Hao Jiang, Vamsi Krishna Ithapu; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 11006-11014

Estimating camera wearer's body pose from an egocentric view (egopose) is a vita l task in augmented and virtual reality. Existing approaches either use a narrow field of view front facing camera that barely captures the wearer, or an extend ed head-mounted top-down camera for maximal wearer visibility. In this paper, we tackle the egopose estimation from a more natural human vision span, where came ra wearer can be seen in the peripheral view and depending on the head pose the wearer may become invisible or has a limited partial view. This is a realistic v isual field for user-centric wearable devices like glasses which have front faci ng wide angle cameras. Existing solutions are not appropriate for this setting, and so, we propose a novel deep learning system taking advantage of both the dyn amic features from camera SLAM and the body shape imagery. We compute 3D head po se, 3D body pose, the figure/ground separation, all at the same time while expli citly enforcing a certain geometric consistency across pose attributes. We furth er show that this system can be trained robustly with lots of existing mocap dat a so we do not have to collect and annotate large new datasets. Lastly, our syst em estimates egopose in real time and on the fly while maintaining high accuracy .

****

Prototypical Matching and Open Set Rejection for Zero-Shot Semantic Segmentation

Hui Zhang, Henghui Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6974-6983

The deep learning methods in addressing semantic segmentation typically demand v

ast amount of pixel-wise annotated training samples. In this work, we present ze
ro-shot semantic segmentation, which aims to identify not only the seen classes
contained in training but also the novel classes that have never been seen. We a
dopt a stringent inductive setting in which only the instances of seen classes a
re accessible during training. We propose an open-aware prototypical matching ap
proach to accomplish the segmentation. The prototypical way extracts the visual
representations by a set of prototypes, making it convenient and flexible to add
 new unseen classes. A prototype projection is trained to map the semantic repre
sentations towards prototypes based on seen instances, and will generate prototy
pes for unseen classes. Moreover, an open-set rejection is utilized to detect th
e objects that do not belong to any seen classes, which greatly reduces the misc
lassifications of unseen objects as seen classes caused by the lack of unseen tr
aining instances. We apply the framework on two segmentation datasets, Pascal VO
C 2012 and Pascal Context, and achieve impressively state-of-the-art performance
.
****

GarmentNets: Category-Level Pose Estimation for Garments via Canonical Space Sha
pe Completion
Cheng Chi, Shuran Song; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2021, pp. 3324-3333
This paper tackles the task of category-level pose estimation for garments. With
 a near infinite degree of freedom, a garment's full configuration (i.e., poses)
 is often described by the per-vertex 3D locations of its entire 3D surface. How
ever, garments are also commonly subject to extreme cases of self-occlusion, esp
ecially when folded or crumpled, making it challenging to perceive their full 3D
 surface. To address these challenges, we propose GarmentNets, where the key ide
a is to formulate the deformable object pose estimation problem as a shape compl
etion task in the canonical space. This canonical space is defined across garmen
ts instances within a category, therefore, specifies the shared category-level p
ose. By mapping the observed partial surface to the canonical space and completi
ng it in this space, the output representation describes the garment's full conf
iguration using a complete 3D mesh with the per-vertex canonical coordinate labe
l. To properly handle the thin 3D structure presented on garments, we proposed a
 novel 3D shape representation using the generalized winding number field. Exper
iments demonstrate that GarmentNets is able to generalize to unseen garment inst
ances and achieve significantly better performance compared to alternative appro
aches. Code and data will be available online.
****

Reliably Fast Adversarial Training via Latent Adversarial Perturbation
Geon Yeong Park, Sang Wan Lee; Proceedings of the IEEE/CVF International Confere
nce on Computer Vision (ICCV), 2021, pp. 7758-7767
While multi-step adversarial training is widely popular as an effective defense
method against strong adversarial attacks, its computational cost is notoriously
 expensive, compared to standard training. Several single-step adversarial train
ing methods have been proposed to mitigate the above-mentioned overhead cost; ho
wever, their performance is not sufficiently reliable depending on the optimizat
ion setting. To overcome such limitations, we deviate from the existing input-sp
ace-based adversarial training regime and propose a single-step latent adversari
al training method (SLAT), which leverages the gradients of latent representatio
n as the latent adversarial perturbation. We demonstrate that the L1 norm of fea
ture gradients is implicitly regularized through the adopted latent perturbation
, thereby recovering local linearity and ensuring reliable performance, compared
 to the existing single-step adversarial training methods. Because latent pertur
bation is based on the gradients of the latent representations which can be obta
ined for free in the process of input gradients computation, the proposed method
 costs roughly the same time as the fast gradient sign method. Experiment result
s demonstrate that the proposed method, despite its structural simplicity, outpe
rforms state-of-the-art accelerated adversarial training methods.
****

R-SLAM: Optimizing Eye Tracking From Rolling Shutter Video of the Retina

Jay Shenoy, James Fong, Jeffrey Tan, Austin Roorda, Ren Ng; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4852-4861
We present a method for optimization-based recovery of eye motion from rolling s
hutter video of the retina. Our approach formulates eye tracking as an optimizat
ion problem that jointly estimates the retina's motion and appearance using conv
ex optimization and a constrained version of gradient descent. By incorporating
the rolling shutter imaging model into the formulation of our joint optimization
, we achieve state-of-the-art accuracy both offline and in real-time. We apply o
ur method to retina video captured with an adaptive optics scanning laser ophtha
lmoscope (AOSLO), demonstrating eye tracking at 1 kHz with accuracies below one
arcminute -- over an order of magnitude higher than conventional eye tracking sy
stems.
****

Inference of Black Hole Fluid-Dynamics From Sparse Interferometric Measurements
Aviad Levis, Daeyoung Lee, Joel A. Tropp, Charles F. Gammie, Katherine L. Bouman
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2021, pp. 2340-2349
We develop an approach to recover the underlying properties of fluid-dynamical p
rocesses from sparse measurements. We are motivated by the task of imaging the s
tochastically evolving environment surrounding black holes, and demonstrate how
flow parameters can be estimated from sparse interferometric measurements used i
n radio astronomical imaging. To model the stochastic flow we use spatio-tempora
l Gaussian Random Fields (GRFs). The high dimensionality of the underlying sourc
e video makes direct representation via a GRF's full covariance matrix intractab
le. In contrast, stochastic partial differential equations are able to capture c
orrelations at multiple scales by specifying only local interaction coefficients
. Our approach estimates the coefficients of a space-time diffusion equation tha
t dictates the stationary statistics of the dynamical process. We analyze our ap
proach on realistic simulations of black hole evolution and demonstrate its adva
ntage over state-of-the-art dynamic black hole imaging techniques.
****

Monocular, One-Stage, Regression of Multiple 3D People
Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, Tao Mei; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11179-11
188
This paper focuses on the regression of multiple 3D people from a single RGB ima
ge. Existing approaches predominantly follow a multi-stage pipeline that first d
etects people in bounding boxes and then independently regresses their 3D body m
eshes. In contrast, we propose to Regress all meshes in a One-stage fashion for
Multiple 3D People (termed ROMP). The approach is conceptually simple, bounding
box-free, and able to learn a per-pixel representation in an end-to-end manner.
Our method simultaneously predicts a Body Center heatmap and a Mesh Parameter ma
p, which can jointly describe the 3D body mesh on the pixel level. Through a bod
y-center-guided sampling process, the body mesh parameters of all people in the
image are easily extracted from the Mesh Parameter map. Equipped with such a fin
e-grained representation, our one-stage framework is free of the complex multi-s
tage process and more robust to occlusion. Compared with state-of-the-art method
s, ROMP achieves superior performance on the challenging multi-person benchmarks
, including 3DPW and CMU Panoptic. Experiments on crowded/occluded datasets demo
nstrate the robustness under various types of occlusion. The code, released at h
ttps://github.com/Arthur151/ROMP, is the first real-time implementation of monoc
ular multi-person 3D mesh regression.
****

PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation
Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu,
Jianping Shi, Lizhuang Ma; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2021, pp. 8761-8770
Cross-domain object detection and semantic segmentation have witnessed impressiv
e progress recently. Existing approaches mainly consider the domain shift result
ing from external environments including the changes of background, illumination

or weather, while distinct camera intrinsic parameters appear commonly in different domains and their influence for domain adaptation has been very rarely explored. In this paper, we observe that the Field of View (FoV) gap induces noticeable instance appearance differences between the source and target domains. We further discover that the FoV gap between two domains impairs domain adaptation performance under both the FoV-increasing (source FoV < target FoV) and FoV-decreasing cases. Motivated by the observations, we propose the Position-Invariant Transform (PIT) to better align images in different domains. We also introduce a reverse PIT for mapping the transformed/aligned images back to the original image space, and design a loss re-weighting strategy to accelerate the training process. Our method can be easily plugged into existing cross-domain detection/segmentation frameworks, while bringing about negligible computational overhead. Extensive experiments demonstrate that our method can soundly boost the performance on both cross-domain object detection and segmentation for state-of-the-art techniques. Our code is available at https://github.com/sheepooo/PIT-Position-Invariant-Transform.

****

Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering

Corentin Dancette, Rémi Cadène, Damien Teney, Matthieu Cord; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1574-1583

We introduce an evaluation methodology for visual question answering (VQA) to better diagnose cases of shortcut learning. These cases happen when a model exploits spurious statistical regularities to produce correct answers but does not actually deploy the desired behavior. There is a need to identify possible shortcuts in a dataset and assess their use before deploying a model in the real world. The research community in VQA has focused exclusively on question-based shortcuts, where a model might, for example, answer "What is the color of the sky" with "blue" by relying mostly on the question-conditional training prior and give little weight to visual evidence. We go a step further and consider multimodal shortcuts that involve both questions and images. We first identify potential shortcuts in the popular VQA v2 training set by mining trivial predictive rules such as co-occurrences of words and visual elements. We then introduce VQA-CounterExamples (VQA-CE), an evaluation protocol based on our subset of CounterExamples i.e. image-question-answer triplets where our rules lead to incorrect answers. We use this new evaluation in a large-scale study of existing approaches for VQA. We demonstrate that even state-of-the-art models perform poorly and that existing techniques to reduce biases are largely ineffective in this context. Our findings suggest that past work on question-based biases in VQA has only addressed one facet of a complex issue. The code for our method is available at \url https://github.com/cdancette/detect-shortcuts

****

H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giró-i-Nieto, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5620-5629

Recent learning approaches that implicitly represent surface geometry using coordinate-based neural representations have shown impressive results in the problem of multi-view 3D reconstruction. The effectiveness of these techniques is, however, subject to the availability of a large number (several tens) of input views of the scene, and computationally demanding optimizations. In this paper, we tackle these limitations for the specific problem of few-shot full 3D head reconstruction, by endowing coordinate-based representations with a probabilistic shape prior that enables faster convergence and better generalization when using few input images (down to three). First, we learn a shape model of 3D heads from thousands of incomplete raw scans using implicit representations. At test time, we jointly overfit two coordinate-based neural networks to the scene, one modeling the geometry and another estimating the surface radiance, using implicit differentiable rendering. We devise a two-stage optimization strategy in which the learned prior is used to initialize and constrain the geometry during an initial opt

imization phase. Then, the prior is unfrozen and fine-tuned to the scene. By doi ng this, we achieve high-fidelity head reconstructions, including hair and shoul ders, and with a high level of detail that consistently outperforms both state-o f-the-art 3D Morphable Models methods in the few-shot scenario, and non-parametr ic methods when large sets of views are available.
****

Image Harmonization With Transformer
Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, Junyu Dong; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 021, pp. 14870-14879
Image harmonization, aiming to make composite images look more realistic, is an important and challenging task. The composite, synthesized by combining foregrou nd from one image with background from another image, inevitably suffers from th e issue of inharmonious appearance caused by distinct imaging conditions, i.e., lights. Current solutions mainly adopt an encoder-decoder architecture with conv olutional neural network (CNN) to capture the context of composite images, tryin g to understand what it looks like in the surrounding background near the foregr ound. In this work, we seek to solve image harmonization with Transformer, by le veraging its powerful ability of modeling long-range context dependencies, for a djusting foreground light to make it compatible with background light while keep ing structure and semantics unchanged. We present the design of our harmonizatio n Transformer frameworks without and with disentanglement, as well as comprehens ive experiments and ablation study, demonstrating the power of Transformer and i nvestigating the Transformer for vision. Our method achieves state-of-the-art pe rformance on both image harmonization and image inpainting/enhancement, indicati ng its superiority. Our code and models are available at https://github.com/zhen glab/HarmonyTransformer.
****

Video Question Answering Using Language-Guided Deep Compressed-Domain Video Feat ure
Nayoung Kim, Seong Jong Ha, Je-Won Kang; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2021, pp. 1708-1717
Video Question Answering (Video QA) aims to give an answer to the question throu gh semantic reasoning between visual and linguistic information. Recently, handl ing large amounts of multi-modal video and language information of a video is co nsidered important in the industry. However, the current video QA models use dee p features, suffered from significant computational complexity and insufficient representation capability both in training and testing. Existing features are ex tracted using pre-trained networks after all the frames are decoded, which is no t always suitable for video QA tasks. In this paper, we develop a novel deep neu ral network to provide video QA features obtained from coded video bit-stream to reduce the complexity. The proposed network includes several dedicated deep mod ules to both the video QA and the video compression system, which is the first a ttempt at the video QA task. The proposed network is predominantly model-agnosti c. It is integrated into the state-of-the-art networks for improved performance without any computationally expensive motion-related deep models. The experiment al results demonstrate that the proposed network outperforms the previous studie s at lower complexity.
****

Human Pose Regression With Residual Log-Likelihood Estimation
Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, Cewu Lu; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 21, pp. 11025-11034
Heatmap-based methods dominate in the field of human pose estimation by modellin g the output distribution through likelihood heatmaps. In contrast, regression-b ased methods are more efficient but suffer from inferior performance. In this wo rk, we explore maximum likelihood estimation (MLE) to develop an efficient and e ffective regression-based method. From the perspective of MLE, adopting differen t regression losses is making different assumptions about the output density fun ction. A density function closer to the true distribution leads to a better regr

ession performance. In light of this, we propose a novel regression paradigm wit h Residual Log-likelihood Estimation (RLE) to capture the underlying output dist ribution. Concretely, RLE learns the change of the distribution instead of the u nreferenced underlying distribution to facilitate the training process. With the proposed reparameterization design, our method is compatible with off-the-shelf flow models. The proposed method is effective, efficient and flexible. We show its potential in various human pose estimation tasks with comprehensive experime nts. Compared to the conventional regression paradigm, regression with RLE bring 12.4 mAP improvement on MSCOCO without any test-time overhead. Moreover, for th e first time, especially on multi-person pose estimation, our regression method is superior to the heatmap-based methods. Our code is available at https://githu b.com/Jeff-sjtu/res-loglikelihood-regression.

****

Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis

Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, Iddo Drori; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 28 6-295

Measuring the acoustic characteristics of a space is often done by capturing its impulse response (IR), a representation of how a full-range stimulus sound exci tes it. This work generates an IR from a single image, which can then be applied to other signals using convolution, simulating the reverberant characteristics of the space shown in the image. Recording these IRs is both time-intensive and expensive, and often infeasible for inaccessible locations. We use an end-to-end neural network architecture to generate plausible audio impulse responses from single images of acoustic environments. We evaluate our method both by compariso ns to ground truth data and by human expert evaluation. We demonstrate our appro ach by generating plausible impulse responses from diverse settings and formats including well known places, musical halls, rooms in paintings, images from anim ations and computer games, synthetic environments generated from text, panoramic images, and video conference backgrounds.

****

Boosting Monocular Depth Estimation With Lightweight 3D Point Fusion

Lam Huynh, Phong Nguyen, Ji■í Matas, Esa Rahtu, Janne Heikkilä; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12767-12776

In this paper, we propose enhancing monocular depth estimation by adding 3D poin ts as depth guidance. Unlike existing depth completion methods, our approach per forms well on extremely sparse and unevenly distributed point clouds, which make s it agnostic to the source of the 3D points. We achieve this by introducing a n ovel multi-scale 3D point fusion network that is both lightweight and efficient. We demonstrate its versatility on two different depth estimation problems where the 3D points have been acquired with conventional structure-from-motion and Li DAR. In both cases, our network performs on par with state-of-the-art depth comp letion methods and achieves significantly higher accuracy when only a small numb er of points is used while being more compact in terms of the number of paramete rs. We show that our method outperforms some contemporary deep learning based mu lti-view stereo and structure-from-motion methods both in accuracy and in compac tness.

****

Spatio-Temporal Dynamic Inference Network for Group Activity Recognition

Hangjie Yuan, Dong Ni, Mang Wang; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2021, pp. 7476-7485

Group activity recognition aims to understand the activity performed by a group of people. In order to solve it, modeling complex spatio-temporal interactions i s the key. Previous methods are limited in reasoning on a predefined graph, whic h ignores the inherent person-specific interaction context. Moreover, they adopt inference schemes that are computationally expensive and easily result in the o ver-smoothing problem. In this paper, we manage to achieve spatio-temporal perso n-specific inferences by proposing Dynamic Inference Network (DIN), which compos es of Dynamic Relation (DR) module and Dynamic Walk (DW) module. We firstly prop

ose to initialize interaction fields on a primary spatio-temporal graph. Within each interaction field, we apply DR to predict the relation matrix and DW to predict the dynamic walk offsets in a joint-processing manner, thus forming a person-specific interaction graph. By updating features on the specific graph, a person can possess a global-level interaction field with a local initialization. Experiments indicate both modules' effectiveness. Moreover, DIN achieves significant improvement compared to previous state-of-the-art methods on two popular datasets under the same setting, while costing much less computation overhead of the reasoning module.

****

Removing the Bias of Integral Pose Regression

Kerui Gu, Linlin Yang, Angela Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11067-11076

Heatmap-based detection methods are dominant for 2D human pose estimation even though regression is more intuitive. The introduction of the integral regression method, which, architecture-wise uses an implicit heatmap, brings the two approaches even closer together. This begs the question -- does detection really outperform regression? In this paper, we investigate the difference in supervision between the heatmap-based detection and integral regression, as this is the key remaining difference between the two approaches. In the process, we discover an underlying bias behind integral pose regression that arises from taking the expectation after the softmax function. To counter the bias, we present a compensation method which we find to improve integral regression accuracy on all 2D pose estimation benchmarks. We further propose a simple joint detection and bias-compensated regression method that considerably outperforms state-of-the-art baselines with few added components.

****

High Quality Disparity Remapping With Two-Stage Warping

Bing Li, Chia-Wen Lin, Cheng Zheng, Shan Liu, Junsong Yuan, Bernard Ghanem, C.-C. Jay Kuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2269-2278

A high quality disparity remapping method that preserves 2D shapes and 3D structures, and adjusts disparities of important objects in stereo image pairs is proposed. It is formulated as a constrained optimization problem, whose solution is challenging, since we need to meet multiple requirements of disparity remapping simultaneously. The one-stage optimization process either degrades the quality of important objects or introduces serious distortions in background regions. To address this challenge, we propose a two-stage warping process to solve it. In the first stage, we develop a warping model that finds the optimal warping grids for important objects to fulfill multiple requirements of disparity remapping. In the second stage, we derive another warping model to refine warping results in less important regions by eliminating serious distortions in shape, disparity and 3D structure. The superior performance of the proposed method is demonstrated by experimental results

****

TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation

Samuel G. Müller, Frank Hutter; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 774-782

Automatic augmentation methods have recently become a crucial pillar for strong model performance in vision tasks. While existing automatic augmentation methods need to trade off simplicity, cost and performance, we present a most simple baseline, TrivialAugment, that outperforms previous methods for almost free. TrivialAugment is parameter-free and only applies a single augmentation to each image. Thus, TrivialAugment's effectiveness is very unexpected to us and we performed very thorough experiments to study its performance. First, we compare TrivialAugment to previous state-of-the-art methods in a variety of image classification scenarios. Then, we perform multiple ablation studies with different augmentation spaces, augmentation methods and setups to understand the crucial requirements for its performance. Additionally, we provide a simple interface to facilitate the widespread adoption of automatic augmentation methods, as well as our full c

ode base for reproducibility. Since our work reveals a stagnation in many parts of automatic augmentation research, we end with a short proposal of best practices for sustained future progress in automatic augmentation methods.
****

Learning To Discover Reflection Symmetry via Polar Matching Convolution
Ahyun Seo, Woohyeon Shim, Minsu Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1285-1294
The task of reflection symmetry detection remains challenging due to significant variations and ambiguities of symmetry patterns in the wild. Furthermore, since the local regions are required to match in reflection for detecting a symmetry pattern, it is hard for standard convolutional networks, which are not equivariant to rotation and reflection, to learn the task. To address the issue, we introduce a new convolutional technique, dubbed the polar matching convolution, which leverages a polar feature pooling, a self-similarity encoding, and a systematic kernel design for axes of different angles. The proposed high-dimensional kernel convolution network effectively learns to discover symmetry patterns from real-world images, overcoming the limitations of standard convolution. In addition, we present a new dataset and introduce a self-supervised learning strategy by augmenting the dataset with synthesizing images. Experiments demonstrate that our method outperforms state-of-the-art methods in terms of accuracy and robustness.
****

Holistic Pose Graph: Modeling Geometric Structure Among Objects in a Scene Using Graph Inference for 3D Object Prediction
Jiwei Xiao, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12717-12726
Due to the missing depth cues, it is essentially ambiguous to detect 3D objects from a single RGB image. Existing methods predict the 3D pose for each object independently or merely by combining local relationships within limited surroundings, but rarely explore the inherent geometric relationships from a global perspective. To address this issue, we argue that modeling geometric structure among objects in a scene is very crucial, and thus elaborately devise the Holistic Pose Graph (HPG) that explicitly integrates all geometric poses including the object pose treated as nodes and the relative pose treated as edges. The inference of the HPG uses GRU to encode the pose features from their corresponding regions in a single RGB image, and passes messages along the graph structure iteratively to improve the predicted poses. To further enhance the correspondence between the object pose and the relative pose, we propose a novel consistency loss to explicitly measure the deviations between them. Finally, we apply Holistic Pose Estimation (HPE) to jointly evaluate both the independent object pose and the relative pose. Our experiments on the SUN RGB-D dataset demonstrate that the proposed method provides a significant improvement on 3D object prediction.
****

Crossover Learning for Fast Online Video Instance Segmentation
Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, Wenyu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8043-8052
Modeling temporal visual context across frames is critical for video instance segmentation (VIS) and other video understanding tasks. In this paper, we propose a fast online VIS model termed CrossVIS. For temporal information modeling in VIS, we present a novel crossover learning scheme that uses the instance feature in the current frame to pixel-wisely localize the same instance in other frames. Different from previous schemes, crossover learning does not require any additional network parameters for feature enhancement. By integrating with the instance segmentation loss, crossover learning enables efficient cross-frame instance-to-pixel relation learning and brings cost-free improvement during inference. Besides, a global balanced instance embedding branch is proposed for better and more stable online instance association. We conduct extensive experiments on three challenging VIS benchmarks, i.e., YouTube-VIS-2019, OVIS, and YouTube-VIS-2021 to evaluate our methods. CrossVIS achieves state-of-the-art online VIS performance and shows a decent trade-off between latency and accuracy. Code is available at

https://github.com/hustvl/CrossVIS.
****
3DIAS: 3D Shape Reconstruction With Implicit Algebraic Surfaces
Mohsen Yavartanoo, Jaeyoung Chung, Reyhaneh Neshatavar, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12446-12455
3D Shape representation has substantial effects on 3D shape reconstruction. Primitive-based representations approximate a 3D shape mainly by a set of simple implicit primitives, but the low geometrical complexity of the primitives limits the shape resolution. Moreover, setting a sufficient number of primitives for an arbitrary shape is challenging. To overcome these issues, we propose a constrained implicit algebraic surface as the primitive with few learnable coefficients and higher geometrical complexities and a deep neural network to produce these primitives. Our experiments demonstrate the superiorities of our method in terms of representation power compared to the state-of-the-art methods in single RGB image 3D shape reconstruction. Furthermore, we show that our method can semantically learn segments of 3D shapes in an unsupervised manner. The code is publicly available from this link.
****
DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection
Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, Chi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8681-8690
Few-shot object detection, which aims at detecting novel objects rapidly from extremely few annotated examples of previously unseen classes, has attracted significant research interest in the community. Most existing approaches employ the Faster R-CNN as basic detection framework, yet, due to the lack of tailored considerations for data-scarce scenario, their performance is often not satisfactory. In this paper, we look closely into the conventional Faster R-CNN and analyze its contradictions from two orthogonal perspectives, namely multi-stage (RPN vs. RCNN) and multi-task (classification vs. localization). To resolve these issues, we propose a simple yet effective architecture, named Decoupled Faster R-CNN (DeFRCN). To be concrete, we extend Faster R-CNN by introducing Gradient Decoupled Layer for multi-stage decoupling and Prototypical Calibration Block for multi-task decoupling. The former is a novel deep layer with redefining the feature-forward operation and gradient-backward operation for decoupling its subsequent layer and preceding layer, and the latter is an offline prototype-based classification model with taking the proposals from detector as input and boosting the original classification scores with additional pairwise scores for calibration. Extensive experiments on multiple benchmarks show our framework is remarkably superior to other existing approaches and establishes a new state-of-the-art in few-shot literature.
****
Multiview Pseudo-Labeling for Semi-Supervised Learning From Video
Bo Xiong, Haoqi Fan, Kristen Grauman, Christoph Feichtenhofer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7209-7219
We present a multiview pseudo-labeling approach to video learning, a novel framework that uses complementary views in the form of appearance and motion information for semi-supervised learning in video. The complementary views help obtain more reliable "pseudo-labels"" on unlabeled video, to learn stronger video representations than from purely supervised data. Though our method capitalizes on multiple views, it nonetheless trains a model that is shared across appearance and motion input and thus, by design, incurs no additional computation overhead at inference time. On multiple video recognition datasets, our method substantially outperforms its supervised counterpart, and compares favorably to previous work on standard benchmarks in self-supervised video representation learning.
****
SketchLattice: Latticed Representation for Sketch Manipulation
Yonggang Qi, Guoyao Su, Pinaki Nath Chowdhury, Mingkang Li, Yi-Zhe Song; Proceed

ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 953-961
The key challenge in designing a sketch representation lies with handling the abstract and iconic nature of sketches. Existing work predominantly utilizes either, (i) a pixelative format that treats sketches as natural images employing off-the-shelf CNN-based networks, or (ii) an elaborately designed vector format that leverages the structural information of drawing orders using sequential RNN-based methods. While the pixelative format lacks intuitive exploitation of structural cues, sketches in vector format are absent in most cases limiting their practical usage. Hence, in this paper, we propose a lattice structured sketch representation that not only removes the bottleneck of requiring vector data but also preserves the structural cues that vector data provides. Essentially, sketch lattice is a set of points sampled from the pixelative format of the sketch using a lattice graph. We show that our lattice structure is particularly amenable to structural changes that largely benefits sketch abstraction modeling for generation tasks. Our lattice representation could be effectively encoded using a graph model, that uses significantly fewer model parameters (13.5 times lesser) than existing state-of-the-art. Extensive experiments demonstrate the effectiveness of sketch lattice for sketch manipulation, including sketch healing and image-to-sketch synthesis.
****

Aligning Latent and Image Spaces To Connect the Unconnectable
Ivan Skorokhodov, Grigorii Sotnikov, Mohamed Elhoseiny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14144-14153
In this work, we develop a method to generate infinite high-resolution images with diverse and complex content. It is based on a perfectly equivariant patch-wise generator with synchronous interpolations in the image and latent spaces. Latent codes, when sampled, are positioned on the coordinate grid, and each pixel is computed from an interpolation of the neighboring codes. We modify the AdaIN mechanism to work in such a setup and train a GAN model to generate images positioned between any two latent vectors. At test time, this allows for generating infinitely large images of diverse scenes that transition naturally from one into another. Apart from that, we introduce LHQ: a new dataset of 90k high-resolution nature landscapes. We test the approach on LHQ, LSUN Tower and LSUN Bridge and outperform the baselines by at least 4 times in terms of quality and diversity of the produced infinite images. The project website is located at https://universome.github.io/alis.
****

End-to-End Trainable Trident Person Search Network Using Adaptive Gradient Propagation
Byeong-Ju Han, Kuhyeun Ko, Jae-Young Sim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 925-933
Person search suffers from the conflicting objectives of commonness and uniqueness between the person detection and re-identification tasks that make the end-to-end training of person search networks difficult. In this paper, we propose a trident network for person search that performs detection, re-identification, and part classification together. We also devise a novel end-to-end training method using adaptive gradient weighting that controls the flow of back-propagated gradients through the re-identification and part classification networks according to the quality of the person detection. The proposed method not only prevents the over-fitting but encourages to exploit fine-grained features by incorporating the part classification branch into the person search framework. Experimental results on the CUHK-SYSU and PRW datasets demonstrate that the proposed method achieves the best performance among the state-of-the-art end-to-end person search methods.
****

HandFoldingNet: A 3D Hand Pose Estimation Network Using Multiscale-Feature Guided Folding of a 2D Hand Skeleton
Wencan Cheng, Jae Hyun Park, Jong Hwan Ko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11260-11269

With increasing applications of 3D hand pose estimation in various human-compute r interaction applications, convolution neural networks (CNNs) based estimation models have been actively explored. However, the existing models require complex architectures or redundant computational resources to trade with the acceptable accuracy. To tackle this limitation, this paper proposes HandFoldingNet, an acc urate and efficient hand pose estimator that regresses the hand joint locations from the normalized 3D hand point cloud input. The proposed model utilizes a fol ding-based decoder that folds a given 2D hand skeleton into the corresponding jo int coordinates. For higher estimation accuracy, folding is guided by multi-scal e features, which include both global and joint-wise local features. Experimenta l results show that the proposed model outperforms the existing methods on three hand pose benchmark datasets with the lowest model parameter requirement. Code is available at https://github.com/cwc1260/HandFold.

****

Learning Deep Local Features With Multiple Dynamic Attentions for Large-Scale Im age Retrieval

Hui Wu, Min Wang, Wengang Zhou, Houqiang Li; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 11416-11425

In image retrieval, learning local features with deep convolutional networks has been demonstrated effective to improve the performance. To discriminate deep lo cal features, some research efforts turn to attention learning. However, existin g attention-based methods only generate a single attention map for each image, w hich limits the exploration of diverse visual patterns. To this end, we propose a novel deep local feature learning architecture to simultaneously focus on mult iple discriminative local patterns in an image. In our framework, we first adap tively reorganize the channels of activation maps for multiple heads. For each h ead, a new dynamic attention module is designed to learn the potential attention s. The whole architecture is trained as metric learning of weighted-sum-pooled g lobal image features, with only image-level relevance label. After the architect ure training, for each database image, we select local features based on their m ulti-head dynamic attentions, which are further indexed for efficient retrieval. Extensive experiments show the proposed method outperforms the state-of-the-art methods on the Revisited Oxford and Paris datasets. Besides, it typically achie ves competitive results even using local features with lower dimensions.

****

Polarimetric Helmholtz Stereopsis

Yuqi Ding, Yu Ji, Mingyuan Zhou, Sing Bing Kang, Jinwei Ye; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5037-5046

Helmholtz stereopsis (HS) exploits the reciprocity principle of light propagatio n (i.e., the Helmholtz reciprocity) for 3D reconstruction of surfaces with arbit rary reflectance. In this paper, we present the polarimetric Helmholtz stereopsi s (polar-HS), which extends the classical HS by considering the polarization sta te of light in the reciprocal paths. With the additional phase information from polarization, polar-HS requires only one reciprocal image pair. We formulate new reciprocity and diffuse/specular polarimetric constraints to recover surface de pths and normals using an optimization framework. Using a hardware prototype, we show that our approach produces high-quality 3D reconstruction for different ty pes of surfaces, ranging from diffuse to highly specular.

****

Motion Prediction Using Trajectory Cues

Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, Meng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 13299-13308

Predicting human motion from a historical pose sequence is at the core of many a pplications in computer vision. Current state-of-the-art methods concentrate on learning motion contexts in the pose space, however, the high dimensionality and complex nature of human pose invoke inherent difficulties in extracting such co ntexts. In this paper, we instead advocate to model motion contexts in the joint trajectory space, as the trajectory of a joint is smooth, vectorial, and gives sufficient information to the model. Moreover, most existing methods consider on

ly the dependencies between skeletal connected joints, disregarding prior knowle dge and the hidden connections between geometrically separated joints. Motivated by this, we present a semi-constrained graph to explicitly encode skeletal conn ections and prior knowledge, while adaptively learn implicit dependencies betwee n joints. We also explore the applications of our approach to a range of objects including human, fish, and mouse. Surprisingly, our method sets the new state-o f-the-art performance on 4 different benchmark datasets, a remarkable highlight is that it achieves a 19.1% accuracy improvement over current state-of-the-art i n average. To facilitate future research, we have released our code at https://g ithub.com/Pose-Group/MPT.

****

Generalized Source-Free Domain Adaptation

Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, Shangling Jui; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8978-8987

Domain adaptation (DA) aims to transfer the knowledge learned from source domain to an unlabeled target domain. Some recent works tackle source-free domain adap tation (SFDA) where only source pre-trained model is available for adaptation to target domain. However those methods does not consider keeping source performan ce which is of high practical value in real world application. In this paper, we propose a new domain adaptation paradigm denoted as Generalized Source-free Dom ain Adaptation (G-SFDA), where the learned model needs to perform well on both t arget and source domains, with only access to current unlabeled target data duri ng adaptation. First, we propose local structure clustering (LSC), aiming to clu ster the target features with its semantically similar neighbors, which successf ully adapts the model to target domain in absence of source data. Second, we pro pose randomly generated domain attention (RGDA), it produces binary domain speci fic attention to activate different feature channels for different domains, mean while the domain attention will be utilized to regularize the gradient during ad aptation to keep source information. In the experiments, for target performance our method is on par with or better than existing DA and SFDA methods, specifica lly achieves state-of-the-art performance (85.4%) on VisDA, and our method works well for all domains after adapting to single or multiple target domains.

****

DisUnknown: Distilling Unknown Factors for Disentanglement Learning

Sitao Xiang, Yuming Gu, Pengda Xiang, Menglei Chai, Hao Li, Yajie Zhao, Mingming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 14810-14819

Disentangling data into interpretable and independent factors is critical for co ntrollable generation tasks. With the availability of labeled data, supervision can help enforce the separation of specific factors as expected. However, it is often expensive or even impossible to label every single factor to achieve fully -supervised disentanglement. In this paper, we adopt a general setting where all factors that are hard to label or identify are encapsulated as a single unknown factor. Under this setting, we propose a flexible weakly-supervised multi-facto r disentanglement framework DisUnknown, which Distills Unknown factors for enabl ing multi-conditional generation regarding both labeled and unknown factors. Spe cifically, a two-stage training approach is adopted to first disentangle the unk nown factor with an effective and robust training method, and then train the fin al generator with the proper disentanglement of all labeled factors utilizing th e unknown distillation. To demonstrate the generalization capacity and scalabili ty of our method, we evaluate it on multiple benchmark datasets qualitatively an d quantitatively and further apply it to various real-world applications on comp licated datasets.

****

Self-Mutating Network for Domain Adaptive Segmentation in Aerial Images

Kyungsu Lee, Haeyun Lee, Jae Youn Hwang; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2021, pp. 7068-7077

The domain-adaptive semantic segmentation in aerial images by a deep-learning te chnique remains a challenge owing to the domain gaps caused by a resolution, ima

ge sensors, time-zone, the density of buildings, and even building styles of eac h city. Currently, convolutional neural network (CNN)-based domain adaptation me thodologies have been developed to decrease the domain gaps, but, they have show n still poor performance to utilize multiple aerial images in different domains. In this paper, therefore, the CNN-based network denoted as Self-Mutating Networ k, which changes the values of parameters of convolutional filters itself accord ing to the domain of input image, is proposed. By adopting Parameter Mutation to change the values of parameters and Parameter Fluctuation to randomly convulse the parameters, the network self-changes and fine-tunes the parameters, then ach ieves better predictions of a domain-adaptive segmentation. Through the ablation study of the Self-Mutating Network, we concluded that the Self-Mutating Network can be utilized in the domain-adaptive semantic segmentation of aerial images i n different domains.

****

3D Building Reconstruction From Monocular Remote Sensing Images

Weijia Li, Lingxuan Meng, Jinwang Wang, Conghui He, Gui-Song Xia, Dahua Lin; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202 1, pp. 12548-12557

3D building reconstruction from monocular remote sensing imagery is an important research problem and an economic solution to large-scale city modeling, compare d with reconstruction from LiDAR data and multi-view imagery. However, several c hallenges such as the partial invisibility of building footprints and facades, t he serious shadow effect, and the extreme variance of building height in large-s cale areas, have restricted the existing monocular image based building reconstr uction studies to certain application scenes, i.e., modeling simple low-rise bui ldings from near-nadir images. In this study, we propose a novel 3D building rec onstruction method for monocular remote sensing images, which tackles the above difficulties, thus providing an appealing solution for more complicated scenario s. We design a multi-task building reconstruction network, named MTBR-Net, to le arn the geometric property of oblique images, the key components of a 3D buildin g model and their relations via four semantic-related and three offset-related t asks. The network outputs are further integrated by a prior knowledge based 3D m odel optimization method to produce the the final 3D building models. Results on a public 3D reconstruction dataset and a novel released dataset demonstrate tha t our method improves the height estimation performance by over 40% and the segm entation F1-score by 2% - 4% compared with current state-of-the-art.

****

Multi-Scale Separable Network for Ultra-High-Definition Video Deblurring

Senyou Deng, Wenqi Ren, Yanyang Yan, Tao Wang, Fenglong Song, Xiaochun Cao; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021 , pp. 14030-14039

Although recent research has witnessed a significant progress on the video deblu rring task, these methods struggle to reconcile inference efficiency and visual quality simultaneously, especially on ultra-high-definition (UHD) videos (e.g., 4K resolution). To address the problem, we propose a novel deep model for fast a nd accurate UHD Video Deblurring (UHDVD). The proposed UHDVD is achieved by a se parable-patch architecture, which collaborates with a multi-scale integration sc heme to achieve a large receptive field without adding the number of generic con volutional layers and kernels. Additionally, we design a residual channel-spatia l attention (RCSA) module to improve accuracy and reduce the depth of the networ k appropriately. The proposed UHDVD is the first real-time deblurring model for 4K videos at 35 fps. To train the proposed model, we build a new dataset compris ed of 4K blurry videos and corresponding sharp frames using three different smar tphones. Comprehensive experimental results show that our network performs favor ably against the state-ofthe-art methods on both the 4K dataset and public bench marks in terms of accuracy, speed, and model size.

****

Baking Neural Radiance Fields for Real-Time View Synthesis

Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, Paul Deb evec; Proceedings of the IEEE/CVF International Conference on Computer Vision (I

CCV), 2021, pp. 5875-5884
Neural volumetric representations such as Neural Radiance Fields (NeRF) have emerged as a compelling technique for learning to represent 3D scenes from images with the goal of rendering photorealistic images of the scene from unobserved viewpoints. However, NeRF's computational requirements are prohibitive for real-time applications: rendering views from a trained NeRF requires querying a multilayer perceptron (MLP) hundreds of times per ray. We present a method to train a NeRF, then precompute and store (i.e. ""bake"") it as a novel representation called a Sparse Neural Radiance Grid (SNeRG) that enables real-time rendering on commodity hardware. To achieve this, we introduce 1) a reformulation of NeRF's architecture, and 2) a sparse voxel grid representation with learned feature vectors. The resulting scene representation retains NeRF's ability to render fine geometric details and view-dependent appearance, is compact (averaging less than 90 MB per scene), and can be rendered in real-time (higher than 30 frames per second on a laptop GPU). Actual screen captures are shown in our video.
****

Parallel Rectangle Flip Attack: A Query-Based Black-Box Attack Against Object Detection
Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, Xiaochun Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7697-7707
Object detection has been widely used in many safety-critical tasks, such as autonomous driving. However, its vulnerability to adversarial examples has not been sufficiently studied, especially under the practical scenario of black-box attacks, where the attacker can only access the query feedback of predicted bounding-boxes and top-1 scores returned by the attacked model. Compared with black-box attack to image classification, there are two main challenges in black-box attack to detection. Firstly, even if one bounding-box is successfully attacked, another sub-optimal bounding-box may be detected near the attacked bounding-box. Secondly, there are multiple bounding-boxes, leading to very high attack cost. To address these challenges, we propose a Parallel Rectangle Flip Attack (PRFA) via random search. Specifically, we generate perturbations in each rectangle patch to avoid sub-optimal detection near the attacked region. Besides, utilizing the observation that adversarial perturbations mainly locate around objects' contours and critical points under white-box attacks, the search space of attacked rectangles is reduced to improve the attack efficiency. Moreover, we develop a parallel mechanism of attacking multiple rectangles simultaneously to further accelerate the attack process. Extensive experiments demonstrate that our method can effectively and efficiently attack various popular object detectors, including anchor-based and anchor-free, and generate transferable adversarial examples.
****

ALADIN: All Layer Adaptive Instance Normalization for Fine-Grained Style Similarity
Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, John Collomosse; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11926-11935
We present ALADIN (All Layer AdaIN); a novel architecture for searching images based on the similarity of their artistic style. Representation learning is critical to visual search, where distance in the learned search embedding reflects image similarity. Learning an embedding that discriminates fine-grained variations in style is hard, due to the difficulty of defining and labelling style. ALADIN takes a weakly supervised approach to learning a representation for fine-grained style similarity of digital artworks, leveraging BAM-FG, a novel large-scale dataset of user generated content groupings gathered from the web. ALADIN sets a new state of the art accuracy for style-based visual search over both coarse labelled style data (BAM) and BAM-FG; a new 2.62 million image dataset of 310,000 fine-grained style groupings also contributed by this work.
****

Visio-Temporal Attention for Multi-Camera Multi-Target Association
Yu-Jhe Li, Xinshuo Weng, Yan Xu, Kris M. Kitani; Proceedings of the IEEE/CVF Int

ernational Conference on Computer Vision (ICCV), 2021, pp. 9834-9844
We address the task of Re-Identification (Re-ID) in multi-target multi-camera (MTMC) tracking where we track multiple pedestrians using multiple overlapping uncalibrated (unknown pose) cameras. Since the videos are temporally synchronized and spatially overlapping, we can see a person from multiple views and associate their trajectory across cameras. In order to find the correct association between pedestrians visible from multiple views during the same time window, we extract a visual feature from a tracklet (sequence of pedestrian images) that encodes its similarity and dissimilarity to all other candidate tracklets. We propose a inter-tracklet (person to person) attention mechanism that learns a representation for a target tracklet while taking into account other tracklets across multiple views. Furthermore, to encode the gait and motion of a person, we introduce second intra-tracklet (person-specific) attention module with position embeddings. This second module employs a transformer encoder to learn a feature from a sequence of features over one tracklet. Experimental results on WILDTRACK and our new dataset `ConstructSite' confirm the superiority of our model over state-of-the-art ReID methods (5% and 10% performance gain respectively) in the context of uncalibrated MTMC tracking. While our model is designed for overlapping cameras, we also obtain state-of-the-art results on two other benchmark datasets (MARS and DukeMTMC) with non-overlapping cameras.
****

A Light Stage on Every Desk
Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, Steven M. Seitz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2420-2429
Every time you sit in front of a TV or monitor, your face is actively illuminated by time-varying patterns of light. This paper proposes to use this time-varying illumination for synthetic relighting of your face with any new illumination condition. In doing so, we take inspiration from the light stage work of Debevec et al. [4], who first demonstrated the ability to relight people captured in a controlled lighting environment. Whereas existing light stages require expensive, room-scale spherical capture gantries and exist in only a few labs in the world, we demonstrate how to acquire useful data from a normal TV or desktop monitor. Instead of subjecting the user to uncomfortable rapidly flashing light patterns, we operate on images of the user watching a YouTube video or other standard content. We train a deep network on images plus monitor patterns of a given user and learn to predict images of that user under any target illumination (monitor pattern). Experimental evaluation shows that our method produces realistic relighting results.
****

Multi-Level Curriculum for Training a Distortion-Aware Barrel Distortion Rectification Model
Kang Liao, Chunyu Lin, Lixin Liao, Yao Zhao, Weiyao Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4389-4398
Barrel distortion rectification aims at removing the radial distortion in a distorted image captured by a wide-angle lens. Previous deep learning methods mainly solve this problem by learning the implicit distortion parameters or the nonlinear rectified mapping function in a direct manner. However, this type of manner results in an indistinct learning process of rectification and thus limits the deep perception of distortion. In this paper, inspired by the curriculum learning, we analyze the barrel distortion rectification task in a progressive and meaningful manner. By considering the relationship among different construction levels in an image, we design a multi-level curriculum that disassembles the rectification task into three levels, structure recovery, semantics embedding, and texture rendering. With the guidance of the curriculum that corresponds to the construction of images, the proposed hierarchical architecture enables a progressive rectification and achieves more accurate results. Moreover, we present a novel distortion-aware pre-training strategy to facilitate the initial learning of neural networks, promoting the model to converge faster and better. Experimental results on the synthesized and real-world distorted image datasets show that the pro

posed approach significantly outperforms other learning methods, both qualitativ
ely and quantitatively.
****

DepthInSpace: Exploitation and Fusion of Multiple Video Frames for Structured-Li
ght Depth Estimation
Mohammad Mahdi Johari, Camilla Carta, François Fleuret; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2021, pp. 6039-6048
We present DepthInSpace, a self-supervised deep-learning method for depth estima
tion using a structured-light camera. The design of this method is motivated by
the commercial use case of embedded depth sensors in nowadays smartphones. We fi
rst propose to use estimated optical flow from ambient information of multiple v
ideo frames as a complementary guide for training a single-frame depth estimatio
n network, helping to preserve edges and reduce over-smoothing issues. Utilizing
 optical flow, we also propose to fuse the data of multiple video frames to get
a more accurate depth map. In particular, fused depth maps are more robust in oc
cluded areas and incur less in flying pixels artifacts. We finally demonstrate t
hat these more precise fused depth maps can be used as self-supervision for fine
-tuning a single-frame depth estimation network to improve its performance. Our
models' effectiveness is evaluated and compared with state-of-the-art models on
both synthetic and our newly introduced real datasets. The implementation code,
training procedure, and both synthetic and captured real datasets are available
at https://www.idiap.ch/paper/depthinspace.
****

GeomNet: A Neural Network Based on Riemannian Geometries of SPD Matrix Space and
 Cholesky Space for 3D Skeleton-Based Interaction Recognition
Xuan Son Nguyen; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2021, pp. 13379-13389
In this paper, we propose a novel method for representation and classification o
f two-person interactions from 3D skeleton sequences. The key idea of our approa
ch is to use Gaussian distributions to capture statistics on Rn and those on the
 space of symmetric positive definite (SPD) matrices. The main challenge is how
to parametrize those distributions. Towards this end, we develop methods for emb
edding Gaussian distributions in matrix groups based on the theory of Lie groups
 and Riemannian symmetric spaces. Our method relies on the Riemannian geometry o
f the underlying manifolds and has the advantage of encoding high-order statisti
cs from 3D joint positions. We show that the proposed method achieves competitiv
e results in two-person interaction recognition on two large-scale benchmarks fo
r 3D human activity understanding.
****

Learning Dynamic Interpolation for Extremely Sparse Light Fields With Wide Basel
ines
Mantang Guo, Jing Jin, Hui Liu, Junhui Hou; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2021, pp. 2450-2459
In this paper, we tackle the problem of dense light field (LF) reconstruction fr
om sparsely-sampled ones with wide baselines and propose a learnable model, name
ly dynamic interpolation, to replace the commonly-used geometry warping operatio
n. Specifically, with the estimated geometric relation between input views, we f
irst construct a lightweight neural network to dynamically learn weights for int
erpolating neighbouring pixels from input views to synthesize each pixel of nove
l views independently. In contrast to the fixed and content-independent weights
employed in the geometry warping operation, the learned interpolation weights im
plicitly incorporate the correspondences between the source and novel views and
adapt to different image content information. Then, we recover the spatial corre
lation between the independently synthesized pixels of each novel view by referr
ing to that of input views using a geometry-based spatial refinement module. We
also constrain the angular correlation between the novel views through a dispari
ty-oriented LF structure loss. Experimental results on LF datasets with wide bas
elines show that the reconstructed LFs achieve much higher PSNR/SSIM and preserv
e the LF parallax structure better than state-of-the-art methods. The source cod
e is publicly available at https://github.com/MantangGuo/DI4SLF.

****
Ultra-High-Definition Image HDR Reconstruction via Collaborative Bilateral Learning
Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, Xiuyi Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4449-4458
Existing single image high dynamic range (HDR) reconstruction attempt to expand the range of luminance. They are not effective to generate plausible textures and colors in the reconstructed results, especially for high-density pixels in ultra-high-definition (UHD) images.To address these problems, we propose a new HDR reconstruction network for UHD images by collaboratively learning color and texture details. First, we propose a dual-path network to extract content and chromatic features at a reduced resolution of the low dynamic range (LDR) input. These two types features are used to fit bilatera-space affine models for real-time HDR reconstruction. To extract the main data structure of the LDR input, we propose to use 3D Tucker decomposition and reconstruction to prevents false edges and noise amplification in the learned bilateral grid. As a result, the high-quality content and chromatic features can be reconstructed capitalized on guided bilateral upsampling. Finally, we fuse these two full-resolution feature maps into the HDR reconstructed results.Our proposed method can achieve real-time processing for UHD image (about 160 fps).Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art HDR reconstruction approaches on public benchmarks and real-world UHD images.
****
Visual Relationship Detection Using Part-and-Sum Transformers With Composite Queries
Qi Dong, Zhuowen Tu, Haofu Liao, Yuting Zhang, Vijay Mahadevan, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3550-3559
Computer vision applications such as visual relationship detection and human object interaction can be formulated as a composite (structured) set detection problem in which both the parts (subject, object, and predicate) and the sum (triplet as a whole) are to be detected in a hierarchical fashion. In this paper, we present a new approach, denoted Part-and-Sum detection Transformer (PST), to perform end-to-end visual composite set detection. Different from existing Transformers in which queries are at a single level, we simultaneously model the joint part and sum hypotheses/interactions with composite queries and attention modules. We explicitly incorporate sum queries to enable better modeling of the part-and-sum relations that are absent in the standard Transformers. Our approach also uses novel tensor-based part queries and vector-based sum queries, and models their joint interaction. We report experiments on two vision tasks, visual relationship detection and human object interaction and demonstrate that PST achieves state of the art results among single-stage models, while nearly matching the results of custom designed two-stage models.
****
SLAMP: Stochastic Latent Appearance and Motion Prediction
Adil Kaan Akan, Erkut Erdem, Aykut Erdem, Fatma Güney; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14728-14737
Motion is an important cue for video prediction and often utilized by separating video content into static and dynamic components. Most of the previous work utilizing motion is deterministic but there are stochastic methods that can model the inherent uncertainty of the future. Existing stochastic models either do not reason about motion explicitly or make limiting assumptions about the static part. In this paper, we reason about appearance and motion in the video stochastically by predicting the future based on the motion history. Explicit reasoning about motion without history already reaches the performance of current stochastic models. The motion history further improves the results by allowing to predict consistent dynamics several frames into the future. Our model performs comparably to the state-of-the-art models on the generic video prediction datasets, however, significantly outperforms them on two challenging real-world autonomous driving datasets with complex motion and dynamic background.

****
Learning To Diversify for Single Domain Generalization
Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, Mahsa Baktashmotlagh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 834-843
Domain generalization (DG) aims to generalize a model trained on multiple source (i.e., training) domains to a distributionally different target (i.e., test) domain. In contrast to the DG setup that strictly requires the availability of multiple source domains, this paper considers a more realistic yet challenging scenario, namely Single Domain Generalization (SDG). In this new setting, there is only one source domain available for training, from which the limited diversity may jeopardize the model generalization on unseen target domains. To tackle this problem, we propose a style-complement module to enhance the generalization power of the model by synthesizing images from diverse distributions that are complementary to the source ones. More specifically, we adopt tractable upper and lower bounds of mutual information (MI) between the generated and source samples and perform the two-step optimization iteratively: (1) by minimizing MI upper bound approximation for each pair, the generated images are forced to diversify from the source samples; (2) subsequently, we maximize the lower bound of MI between the samples from the same semantic category, which assists the network to learn discriminative features from diverse-styled images. Extensive experiments on three benchmark datasets demonstrate the superiority of our approach, which surpasses the state-of-the-art single DG methods by up to 25.14%.
****
CPF: Learning a Contact Potential Field To Model the Hand-Object Interaction
Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11097-11106
Modeling the hand-object (HO) interaction not only requires estimation of the HO pose, but also pays attention to the contact due to their interaction. Significant progress has been made in estimating hand and object separately with deep learning methods, simultaneous HO pose estimation and contact modeling has not yet been fully explored. In this paper, we present an explicit contact representation namely Contact Potential Field (CPF), and a learning-fitting hybrid framework namely MIHO to Modeling the Interaction of Hand and Object. In CPF, we treat each contacting HO vertex pair as a spring-mass system. Hence the whole system forms a potential field with minimal elastic energy at the grasp position. Extensive experiments on the two commonly used benchmarks have demonstrated that our method can achieve state-of-the-art in several reconstruction metrics, and allow us to produce more physically plausible HO pose even when the ground-truth exhibits severe interpenetration or disjointedness. Our code is available at https://github.com/lixiny/CPF.
****
Sensor-Guided Optical Flow
Matteo Poggi, Filippo Aleotti, Stefano Mattoccia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7908-7918
This paper proposes a framework to guide an optical flow network with external cues to achieve superior accuracy either on known or unseen domains. Given the availability of sparse yet accurate optical flow hints from an external source, these are injected to modulate the correlation scores computed by a state-of-the-art optical flow network and guide it towards more accurate predictions. Although no real sensor can provide sparse flow hints, we show how these can be obtained by combining depth measurements from active sensors with geometry and hand-crafted optical flow algorithms, leading to accurate enough hints for our purpose. Experimental results with a state-of-the-art flow network on standard benchmarks support the effectiveness of our framework, both in simulated and real conditions.
****
Wasserstein Coupled Graph Learning for Cross-Modal Retrieval
Yun Wang, Tong Zhang, Xueya Zhang, Zhen Cui, Yuge Huang, Pengcheng Shen, Shaoxin

Li, Jian Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1813-1822

Graphs play an important role in cross-modal image-text understanding as they characterize the intrinsic structure which is robust and crucial for the measurement of cross-modal similarity. In this work, we propose a Wasserstein Coupled Graph Learning (WCGL) method to deal with the cross-modal retrieval task. First, graphs are constructed according to two input cross-modal samples separately, and passed through the corresponding graph encoders to extract robust features. Then, a Wasserstein coupled dictionary, containing multiple pairs of counterpart graph keys with each key corresponding to one modality, is constructed for further feature learning. Based on this dictionary, the input graphs can be transformed into the dictionary space to facilitate the similarity measurement through a Wasserstein Graph Embedding (WGE) process. The WGE could capture the graph correlation between the input and each corresponding key through optimal transport, and hence well characterize the inter-graph structural relationship. To further achieve discriminant graph learning, we specifically define a Wasserstein discriminant loss on the coupled graph keys to make the intra-class (counterpart) keys more compact and inter-class (non-counterpart) keys more dispersed, which further promotes the final cross-modal retrieval task. Experimental results demonstrate the effectiveness and state-of-the-art performance.

****

ADNet: Leveraging Error-Bias Towards Normal Direction in Face Alignment

Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, Fangyun Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3080-3090

The recent progress of CNN has dramatically improved face alignment performance. However, few works have paid attention to the error-bias with respect to error distribution of facial landmarks. In this paper, we investigate the error-bias issue in face alignment, where the distributions of landmark errors tend to spread along the tangent line to landmark curves. This error-bias is not trivial since it is closely connected to the ambiguous landmark labeling task. Inspired by this observation, we seek a way to leverage the error-bias property for better convergence of CNN model. To this end, we propose anisotropic direction loss (ADL) and anisotropic attention module (AAM) for coordinate and heatmap regression, respectively. ADL imposes strong binding force in normal direction for each landmark point on facial boundaries. On the other hand, AAM is an attention module which can get anisotropic attention mask focusing on the region of point and its local edge connected by adjacent points, it has a stronger response in tangent than in normal, which means relaxed constraints in the tangent. These two methods work in a complementary manner to learn both facial structures and texture details. Finally, we integrate them into an optimized end-to-end training pipeline named ADNet. Our ADNet achieves state-of-the-art results on 300W, WFLW and COFW datasets, which demonstrates the effectiveness and robustness.

****

Pixel-Perfect Structure-From-Motion With Featuremetric Refinement

Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5987-5997

Finding local features that are repeatable across multiple views is a cornerstone of sparse 3D reconstruction. The classical image matching paradigm detects keypoints per-image once and for all, which can yield poorly-localized features and propagate large errors to the final geometry. In this paper, we refine two key steps of structure-from-motion by a direct alignment of low-level image information from multiple views: we first adjust the initial keypoint locations prior to any geometric estimation, and subsequently refine points and camera poses as a post-processing. This refinement is robust to large detection noise and appearance changes, as it optimizes a featuremetric error based on dense features predicted by a neural network. This significantly improves the accuracy of camera poses and scene geometry for a wide range of keypoint detectors, challenging viewing conditions, and off-the-shelf deep features. Our system easily scales to large image collections, enabling pixel-perfect crowd-sourced localization at scale. O

ur code is publicly available at https://github.com/cvg/pixel-perfect-sfm as an add-on to the popular SfM software COLMAP.
****

BiaSwap: Removing Dataset Bias With Bias-Tailored Swapping Augmentation
Eungyeup Kim, Jihyeon Lee, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14992-15001
Deep neural networks often make decisions based on the spurious correlations inherent in the dataset, failing to generalize in an unbiased data distribution. Although previous approaches pre-define the type of dataset bias to prevent the network from learning it, recognizing the bias type in the real dataset is often prohibitive. This paper proposes a novel bias-tailored augmentation-based approach, BiaSwap, for learning debiased representation without requiring supervision on the bias type. Motivated by the phenomenon that the bias corresponds to the attributes the model learns as a shortcut, we utilize an image-to-image translation model optimized to transfer the attributes that the classifier often learns easily. As a prerequisite, we sort the training samples based on how much a biased model exploits them as a shortcut and divide them into bias-guiding and bias-contrary samples in an unsupervised manner. Afterwards, we utilize the CAM of GCE-trained classifier in the patch cooccurrence discriminator in order to focus on translating the bias attributes. Therefore, given the pair of bias-guiding and bias-contrary, the model generates the augmented bias-swapped image which contains the bias attributes from the bias-contrary images, while preserving bias-irrelevant ones in the bias-guiding images. We demonstrate the superiority of our approach against the baselines over both synthetic and real-world datasets. Even without careful supervision on the bias, BiaSwap achieves a remarkable performance on both unbiased and bias-guiding samples, implying the improved generalization capability of the model.
****

GistNet: A Geometric Structure Transfer Network for Long-Tailed Recognition
Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, Nuno Vasconcelos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8209-8218
The problem of long-tailed recognition, where the number of examples per class is highly unbalanced, is considered. It is hypothesized that the well known tendency of standard classifier training to overfit to popular classes can be exploited for effective transfer learning. Rather than eliminating this overfitting, e.g. by adopting popular class-balanced sampling methods, the learning algorithm should instead leverage this overfitting to transfer geometric information from popular to low-shot classes. A new classifier architecture, GistNet, is proposed to support this goal, using constellations of classifier parameters to encode the class geometry. A new learning algorithm is then proposed for GeometrIc Structure Transfer (GIST), with resort to a combination of loss functions that combine class-balanced and random sampling to guarantee that, while overfitting to the popular classes is restricted to geometric parameters, it is leveraged to transfer class geometry from popular to few-shot classes. This enables better generalization for few-shot classes without the need for the manual specification of class weights, or even the explicit grouping of classes into different types. Experiments on two popular long-tailed recognition datasets show that GistNet outperforms existing solutions to this problem.
****

Distance-Aware Quantization
Dohyung Kim, Junghyup Lee, Bumsub Ham; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5271-5280
We address the problem of network quantization, that is, reducing bit-widths of weights and/or activations to lighten network architectures. Quantization methods use a rounding function to map full-precision values to the nearest quantized ones, but this operation is not differentiable. There are mainly two approaches to training quantized networks with gradient-based optimizers. First, a straight-through estimator (STE) replaces the zero derivative of the rounding with that of an identity function, which causes a gradient mismatch problem. Second, soft quantizers approximate the rounding with continuous functions at training time,

and exploit the rounding for quantization at test time. This alleviates the gradient mismatch, but causes a quantizer gap problem. We alleviate both problems in a unified framework. To this end, we introduce a novel quantizer, dubbed a distance-aware quantizer (DAQ), that mainly consists of a distance-aware soft rounding (DASR) and a temperature controller. To alleviate the gradient mismatch problem, DASR approximates the discrete rounding with the kernel soft argmax, which is based on our insight that the quantization can be formulated as a distance-based assignment problem between full-precision values and quantized ones. The controller adjusts the temperature parameter in DASR adaptively according to the input, addressing the quantizer gap problem. Experimental results on standard benchmarks show that DAQ outperforms the state of the art significantly for various bit-widths without bells and whistles.

****

Shape-Biased Domain Generalization via Shock Graph Embeddings

Maruthi Narayanan, Vickram Rajendran, Benjamin Kimia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1315-1325

There is an emerging sense that the vulnerability of Image Convolutional Neural Networks (CNN), i.e., sensitivity to image corruptions, perturbations, and adversarial attacks, is connected with Texture Bias. This relative lack of Shape Bias is also responsible for poor performance in Domain Generalization (DG). The inclusion of a role of shape alleviates these vulnerabilities and some approaches have achieved this by training on negative images, images endowed with edge maps, or images with conflicting shape and texture information. This paper advocates an explicit and complete representation of shape using a classical computer vision approach, namely, representing the shape content of an image with the shock graph of its contour map. The resulting graph and its descriptor is a complete representation of contour content and is classified using recent Graph Neural Network (GNN) methods. The experimental results on three domain shift datasets, Colored MNIST, PACS, and VLCS demonstrate that even without using appearance the shape-based approach exceeds classical Image CNN based methods in domain generalization.

****

PixelSynth: Generating a 3D-Consistent Experience From a Single Image

Chris Rockwell, David F. Fouhey, Justin Johnson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14104-14113

Recent advancements in differentiable rendering and 3D reasoning have driven exciting results in novel view synthesis from a single image. Despite realistic results, methods are limited to relatively small view change. In order to synthesize immersive scenes, models must also be able to extrapolate. We present an approach that fuses 3D reasoning with autoregressive modeling to outpaint large view changes in a 3D-consistent manner, which enables scene synthesis. We demonstrate considerable improvement in single-image large-angle view synthesis results compared to a variety of methods and possible variants across simulated and real datasets. In addition, we show increased 3D consistency compared to alternative accumulation methods.

****

Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, Christian Theobalt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12959-12970

We present Non-Rigid Neural Radiance Fields (NR-NeRF), a reconstruction and novel view synthesis approach for general non-rigid dynamic scenes. Our approach takes RGB images of a dynamic scene as input (e.g., from a monocular video recording), and creates a high-quality space-time geometry and appearance representation. We show that a single handheld consumer-grade camera is sufficient to synthesize sophisticated renderings of a dynamic scene from novel virtual camera views, e.g. a `bullet-time' video effect. NR-NeRF disentangles the dynamic scene into a canonical volume and its deformation. Scene deformation is implemented as ray bending, where straight rays are deformed non-rigidly. We also propose a novel ri

gidity network to better constrain rigid regions of the scene, leading to more s
table results. The ray bending and rigidity network are trained without explicit
 supervision. Our formulation enables dense correspondence estimation across vie
ws and time, and compelling video editing applications such as motion exaggerati
on. Our code will be open sourced.
****

Learning To Cut by Watching Movies
Alejandro Pardo, Fabian Caba, Juan Léon Alcázar, Ali K. Thabet, Bernard Ghanem;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2021, pp. 6858-6868
Video content creation keeps growing at an incredible pace; yet, creating engagi
ng stories remains challenging and requires non-trivial video editing expertise.
 Many video editing components are astonishingly hard to automate primarily due
to the lack of raw video materials. This paper focuses on a new task for computa
tional video editing, namely the task of raking cut plausibility. Our key idea i
s to leverage content that has already been edited to learn fine-grained audiovi
sual patterns that trigger cuts. To do this, we first collected a data source of
 more than 10K videos, from which we extract more than 260K cuts. We devise a mo
del that learns to discriminate between real and artificial cuts via contrastive
 learning. We set up a new task and a set of baselines to benchmark video cut ge
neration. We observe that our proposed model outperforms the baselines by large
margins. To demonstrate our model in real-world applications, we conduct human s
tudies in a collection of unedited videos. The results show that our model does
a better job at cutting than random and alternative baselines.
****

Sketch2Mesh: Reconstructing and Editing 3D Shapes From Sketches
Benoit Guillard, Edoardo Remelli, Pierre Yvernay, Pascal Fua; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13023-13
032
Reconstructing 3D shape from 2D sketches has long been an open problem because t
he sketches only provide very sparse and ambiguous information. In this paper, w
e use an encoder/decoder architecture for the sketch to mesh translation. When i
ntegrated into a user interface that provides camera parameters for the sketches
, this enables us to leverage its latent parametrization to represent and refine
 a 3D mesh so that its projections match the external contours outlined in the s
ketch. We will show that this approach is easy to deploy, robust to style change
s, and effective. Furthermore, it can be used for shape refinement given only si
ngle pen strokes. We compare our approach to state-of-the-art methods on sketche
s - both hand-drawn and synthesized - and demonstrate that we outperform them.
****

Generic Event Boundary Detection: A Benchmark for Event Segmentation
Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, Matt Feiszli;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2021, pp. 8075-8084
This paper presents a novel task together with a new benchmark for detecting gen
eric, taxonomy-free event boundaries that segment a whole video into chunks. Con
ventional work in temporal video segmentation and action detection focuses on lo
calizing pre-defined action categories and thus does not scale to generic videos
. Cognitive Science has known since last century that humans consistently segmen
t videos into meaningful temporal chunks. This segmentation happens naturally, w
ithout pre-defined event categories and without being explicitly asked to do so.
 Here, we repeat these cognitive experiments on mainstream CV datasets; with our
 novel annotation guideline which addresses the complexities of taxonomy-free ev
ent boundary annotation, we introduce the task of Generic Event Boundary Detecti
on (GEBD) and the new benchmark Kinetics-GEBD. We view GEBD as an important step
ping stone towards understanding the video as a whole, and believe it has been p
reviously neglected due to a lack of proper task definition and annotations. Thr
ough experiment and human study we demonstrate the value of the annotations. Fur
ther, we benchmark supervised and un-supervised GEBD approaches on the TAPOS dat
aset and our Kinetics-GEBD. We release our annotations and baseline codes at CVP

R'21 LOVEU Challenge: https://sites.google.com/view/loveucvpr21.
****
How Shift Equivariance Impacts Metric Learning for Instance Segmentation
Josef Lorenz Rumberger, Xiaoyan Yu, Peter Hirsch, Melanie Dohmen, Vanessa Emanuela Guarino, Ashkan Mokarian, Lisa Mais, Jan Funke, Dagmar Kainmüller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7128-7136
Metric learning has received conflicting assessments concerning its suitability for solving instance segmentation tasks. It has been dismissed as theoretically flawed due to the shift equivariance of the employed CNNs and their respective inability to distinguish same-looking objects. Yet it has been shown to yield state of the art results for a variety of tasks, and practical issues have mainly been reported in the context of tile-and-stitch approaches, where discontinuities at tile boundaries have been observed. To date, neither of the reported issues have undergone thorough formal analysis. In our work, we contribute a comprehensive formal analysis of the shift equivariance properties of encoder-decoder-style CNNs, which yields a clear picture of what can and cannot be achieved with metric learning in the face of same-looking objects. In particular, we prove that a standard encoder-decoder network that takes d-dimensional images as input, with l pooling layers and pooling factor f, has the capacity to distinguish at most f^(dl) same-looking objects, and we show that this upper limit can be reached. Furthermore, we show that to avoid discontinuities in a tile-and-stitch approach, assuming standard batch size 1, it is necessary to employ valid convolutions in combination with a training output window size strictly greater than f^l, while at test-time it is necessary to crop tiles to size n * f^l before stitching, with n >= 1. We complement these theoretical findings by discussing a number of insightful special cases for which we show empirical results on synthetic and real data.
****
Calibrated Adversarial Refinement for Stochastic Semantic Segmentation
Elias Kassapis, Georgi Dikov, Deepak K. Gupta, Cedric Nugteren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7057-7067
In semantic segmentation tasks, input images can often have more than one plausible interpretation, thus allowing for multiple valid labels. To capture such ambiguities, recent work has explored the use of probabilistic networks that can learn a distribution over predictions. However, these do not necessarily represent the empirical distribution accurately. In this work, we present a strategy for learning a calibrated predictive distribution over semantic maps, where the probability associated with each prediction reflects its ground truth correctness likelihood. To this end, we propose a novel two-stage, cascaded approach for calibrated adversarial refinement: (i) a standard segmentation network is trained with categorical cross-entropy to predict a pixelwise probability distribution over semantic classes and (ii) an adversarially trained stochastic network is used to model the inter-pixel correlations to refine the output of the first network into coherent samples. Importantly, to calibrate the refinement network and prevent mode collapse, the expectation of the samples in the second stage is matched to the probabilities predicted in the first. We demonstrate the versatility and robustness of the approach by achieving state-of-the-art results on the multigrader LIDC dataset and on a modified Cityscapes dataset with injected ambiguities. In addition, we show that the core design can be adapted to other tasks requiring learning a calibrated predictive distribution by experimenting on a toy regression dataset. We provide an open source implementation of our method at https://github.com/EliasKassapis/CARSSS.
****
Self-Supervised Visual Representations Learning by Contrastive Mask Prediction
Yucheng Zhao, Guangting Wang, Chong Luo, Wenjun Zeng, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10160-10169
Advanced self-supervised visual representation learning methods rely on the inst

ance discrimination (ID) pretext task. We point out that the ID task has an impl
icit semantic consistency (SC) assumption, which may not hold in unconstrained d
atasets. In this paper, we propose a novel contrastive mask prediction (CMP) tas
k for visual representation learning and design a mask contrast (MaskCo) framewo
rk to implement the idea. MaskCo contrasts region-level features instead of view
-level features, which makes it possible to identify the positive sample without
 any assumptions. To solve the domain gap between masked and unmasked features,
we design a dedicated mask prediction head in MaskCo. This module is shown to be
 the key to the success of the CMP. We evaluated MaskCo on training datasets bey
ond ImageNet and compare its performance with MoCo V2. Results show that MaskCo
achieves comparable performance with MoCo V2 using ImageNet training dataset, bu
t demonstrates a stronger performance across a range of downstream tasks when CO
CO or Conceptual Captions are used for training. MaskCo provides a promising alt
ernative to the ID-based methods for self-supervised learning in the wild.
****

Personalized Image Semantic Segmentation
Yu Zhang, Chang-Bin Zhang, Peng-Tao Jiang, Ming-Ming Cheng, Feng Mao; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
10549-10559
Semantic segmentation models trained on public datasets have achieved great succ
ess in recent years. However, these models didn't consider the personalization i
ssue of segmentation though it is important in practice. In this paper, we addre
ss the problem of personalized image segmentation. The objective is to generate
more accurate segmentation results on unlabeled personalized images by investiga
ting the data's personalized traits. To open up future research in this area, we
 collect a large dataset containing various users' personalized images called PS
S (Personalized Semantic Segmentation). We also survey some recent researches re
lated to this problem and report their performance on our dataset. Furthermore,
by observing the correlation among a user's personalized images, we propose a ba
seline method that incorporates the inter-image context when segmenting certain
images. Extensive experiments show that our method outperforms the existing meth
ods on the proposed dataset. The code and the PSS dataset are available at https
://mmcheng.net/pss/.
****

Fooling LiDAR Perception via Adversarial Trajectory Perturbation
Yiming Li, Congcong Wen, Felix Juefei-Xu, Chen Feng; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 7898-7907
LiDAR point clouds collected from a moving vehicle are functions of its trajecto
ries, because the sensor motion needs to be compensated to avoid distortions. Wh
en autonomous vehicles are sending LiDAR point clouds to deep networks for perce
ption and planning, could the motion compensation consequently become a wide-ope
n backdoor in those networks, due to both the adversarial vulnerability of deep
learning and GPS-based vehicle trajectory estimation that is susceptible to wire
less spoofing? We demonstrate such possibilities for the first time: instead of
directly attacking point cloud coordinates which requires tampering with the raw
 LiDAR readings, only adversarial spoofing of a self-driving car's trajectory wi
th small perturbations is enough to make safety-critical objects undetectable or
 detected with incorrect positions. Moreover, polynomial trajectory perturbation
 is developed to achieve a temporally-smooth and highly-imperceptible attack. Ex
tensive experiments on 3D object detection have shown that such attacks not only
 lower the performance of the state-of-the-art detectors effectively, but also t
ransfer to other detectors, raising a red flag for the community. The code is av
ailable on https://ai4ce.github.io/FLAT/.
****

Extreme Structure From Motion for Indoor Panoramas Without Visual Overlaps
Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, Yasutaka
Furukawa; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2021, pp. 5703-5711
This paper proposes an extreme structure from motion (SfM) algorithm for residen
tial indoor panoramas that have little to no visual overlaps. Only a single pano

rama is present in a room for many cases, making the task infeasible for existing SfM algorithms. Our idea is to learn to evaluate the realism of room/door/window arrangements in the top-down semantic space. After using heuristics to enumerate possible arrangements based on door detections, we evaluate their realism scores, pick the most realistic arrangement, and return the corresponding camera poses. We evaluate the proposed approach on a dataset of 1029 panorama images with 286 houses. Our qualitative and quantitative evaluations show that an existing SfM approach completely fails for most of the houses. The proposed approach achieves the mean positional error of less than 1.0 meter for 47% of the houses and even 78% when considering the top five reconstructions. We will share the code and data in https://github.com/aminshabani/extreme-indoor-sfm.

****

Dynamic Context-Sensitive Filtering Network for Video Salient Object Detection

Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, Zhongxuan Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1553-1563

The ability to capture inter-frame dynamics has been critical to the development of video salient object detection (VSOD). While many works have achieved great success in this field, a deeper insight into its dynamic nature should be developed. In this work, we aim to answer the following questions: How can a model adjust itself to dynamic variations as well as perceive fine differences in the real-world environment; How are the temporal dynamics well introduced into spatial information over time? To this end, we propose a dynamic context-sensitive filtering network (DCFNet) equipped with a dynamic context-sensitive filtering module (DCFM) and an effective bidirectional dynamic fusion strategy. The proposed DCFM sheds new light on dynamic filter generation by extracting location-related affinities between consecutive frames. Our bidirectional dynamic fusion strategy encourages the interaction of spatial and temporal information in a dynamic manner. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance on most VSOD datasets while ensuring a real-time speed of 28 fps. The source code is publicly available at https://github.com/OIPLab-DUT/DCFNet.

****

Boosting the Generalization Capability in Cross-Domain Few-Shot Learning via Noise-Enhanced Supervised Autoencoder

Hanwen Liang, Qiong Zhang, Peng Dai, Juwei Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9424-9434

State of the art (SOTA) few-shot learning (FSL) methods suffer significant performance drop in the presence of domain differences between source and target datasets. The strong discrimination ability on the source dataset does not necessarily translate to high classification accuracy on the target dataset. In this work, we address this cross-domain few-shot learning (CDFSL) problem by boosting the generalization capability of the model. Specifically, we teach the model to capture broader variations of the feature distributions with a novel noise-enhanced supervised autoencoder (NSAE). NSAE trains the model by jointly reconstructing inputs and predicting the labels of inputs as well as their reconstructed pairs. Theoretical analysis based on intra-class correlation (ICC) shows that the feature embeddings learned from NSAE have stronger discrimination and generalization abilities in the target domain. We also take advantage of NSAE structure and propose a two-step fine-tuning procedure that achieves better adaption and improves classification performance in the target domain. Extensive experiments and ablation studies are conducted to demonstrate the effectiveness of the proposed method. Experimental results show that our proposed method consistently outperforms SOTA methods under various conditions.

****

Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, Jamie Shotton; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3681-3691

We demonstrate that it is possible to perform face-related computer vision in th

e wild using synthetic data alone. The community has long enjoyed the benefits o f synthesizing training data with graphics, but the domain gap between real and synthetic data has remained a problem, especially for human faces. Researchers h ave tried to bridge this gap with data mixing, domain adaptation, and domain-adv ersarial training, but we show that it is possible to synthesize data with minim al domain gap, so that models trained on synthetic data generalize to real in-th e-wild datasets. We describe how to combine a procedurally-generated parametric 3D face model with a comprehensive library of hand-crafted assets to render trai ning images with unprecedented realism and diversity. We train machine learning systems for face-related tasks such as landmark localization and face parsing, s howing that synthetic data can both match real data in accuracy, as well as open up new approaches where manual labeling would be impossible.

****

StereOBJ-1M: Large-Scale Stereo Image Dataset for 6D Object Pose Estimation

Xingyu Liu, Shun Iwase, Kris M. Kitani; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2021, pp. 10870-10879

We present a large-scale stereo RGB image object pose estimation dataset named t he StereOBJ-1M dataset. The dataset is designed to address challenging cases suc h as object transparency, translucency, and specular reflection, in addition to the common challenges of occlusion, symmetry, and variations in illumination and environments. In order to collect data of sufficient scale for modern deep lear ning models, we propose a novel method for efficiently annotating pose data in a multi-view fashion that allows data capturing in complex and flexible environme nts. Fully annotated with 6D object poses, our dataset contains over 396K frames and over 1.5M annotations of 18 objects recorded in 183 scenes constructed in 1 1 different environments. The 18 objects include 8 symmetric objects, 7 transpar ent objects, and 8 reflective objects. We benchmark two state-of-the-art pose es timation frameworks on StereOBJ-1M as baselines for future work. We also propose a novel object-level pose optimization method for computing 6D pose from keypoi nt predictions in multiple images.

****

Predictive Feature Learning for Future Segmentation Prediction

Zihang Lin, Jiangxin Sun, Jian-Fang Hu, Qizhi Yu, Jian-Huang Lai, Wei-Shi Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7365-7374

Future segmentation prediction aims to predict the segmentation masks for unobse rved future frames. Most existing works addressed it by directly predicting the intermediate features extracted by existing segmentation models. However, these segmentation features are learned to be local discriminative (with rich details) and are always of high resolution/dimension. Hence, the complicated spatio-temp oral variations of these features are difficult to predict, which motivates us t o learn a more predictive representation. In this work, we develop a novel frame work called Predictive Feature Autoencoder. In the proposed framework, we constr uct an autoencoder which serves as a bridge between the segmentation features an d the predictor. In the latent feature learned by the autoencoder, global struct ures are enhanced and local details are suppressed so that it is more predictive . In order to reduce the risk of vanishing the suppressed details during recurre nt feature prediction, we further introduce a reconstruction constraint in the p rediction module. Extensive experiments show the effectiveness of the proposed a pproach and our method outperforms state-of-the-arts by a considerable margin.

****

PIAP-DF: Pixel-Interested and Anti Person-Specific Facial Action Unit Detection Net With Discrete Feedback Learning

Yang Tang, Wangding Zeng, Dafei Zhao, Honggang Zhang; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 12899-12908

Facial Action Units (AUs) are of great significance in communication. Automatic AU detection can improve the understanding of psychological conditions and emoti onal status. Recently, several deep learning methods have been proposed to detec t AUs automatically. However, several challenges, such as poor extraction of fin e-grained and robust local AUs information, model overfitting on person-specific

features, as well as the limitation of datasets with wrong labels, remain to be addressed. In this paper, we propose a joint strategy called PIAP-DF to solve these problems, which involves 1) a multi-stage Pixel-Interested learning method with pixel-level attention for each AU; 2) an Anti Person-Specific method aiming to eliminate features associated with any individual as much as possible; 3) a semi-supervised learning method with Discrete Feedback, designed to effectively utilize unlabeled data and mitigate the negative impacts of wrong labels. Experimental results on the two popular AU detection datasets BP4D and DISFA prove that PIAP-DF can be the new state-of-the-art method. Compared with the current best method, PIAP-DF improves the average F1 score by 3.2% on BP4D and by 0.5% on DISFA. All modules of PIAP-DF can be easily removed after training to obtain a lightweight model for practical application.

****

NPMs: Neural Parametric Models for 3D Deformable Shapes

Pablo Palafox, Aljaž Boži■, Justus Thies, Matthias Nießner, Angela Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12695-12705

Parametric 3D models have enabled a wide variety of tasks in computer graphics and vision, such as modeling human bodies, faces, and hands. However, the construction of these parametric models is often tedious, as it requires heavy manual tweaking, and they struggle to represent additional complexity and details such as wrinkles or clothing. To this end, we propose Neural Parametric Models (NPMs), a novel, learned alternative to traditional, parametric 3D models, which does not require hand-crafted, object-specific constraints. In particular, we learn to disentangle 4D dynamics into latent-space representations of shape and pose, leveraging the flexibility of recent developments in learned implicit functions. Crucially, once learned, our neural parametric models of shape and pose enable optimization over the learned spaces to fit to new observations, similar to the fitting of a traditional parametric model, e.g., SMPL. This enables NPMs to achieve a significantly more accurate and detailed representation of observed deformable sequences. We show that NPMs improve notably over both parametric and non-parametric state of the art in reconstruction and tracking of monocular depth sequences of clothed humans and hands. Latent-space interpolation as well as shape / pose transfer experiments further demonstrate the usefulness of NPMs. Code is publicly available at https://pablopalafox.github.io/npms.

****

Semantic Aware Data Augmentation for Cell Nuclei Microscopical Images With Artificial Neural Networks

Alireza Naghizadeh, Hongye Xu, Mohab Mohamed, Dimitris N. Metaxas, Dongfang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3952-3961

There exists many powerful architectures for object detection and semantic segmentation of both biomedical and natural images. However, a difficulty arises in the ability to create training datasets that are large and well-varied. The importance of this subject is nested in the amount of training data that artificial neural networks need to accurately identify and segment objects in images and the infeasibility of acquiring a sufficient dataset within the biomedical field. This paper introduces a new data augmentation method that generates artificial cell nuclei microscopical images along with their correct semantic segmentation labels. Data augmentation provides a step toward accessing higher generalization capabilities of artificial neural networks. An initial set of segmentation objects is used with Greedy AutoAugment to find the strongest performing augmentation policies. The found policies and the initial set of segmentation objects are then used in the creation of the final artificial images. When comparing the state-of-the-art data augmentation methods with the proposed method, the proposed method is shown to consistently outperform current solutions in the generation of nuclei microscopical images.

****

NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-View Stereo

Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5610-5619

In this work, we present a new multi-view depth estimation method that utilizes both conventional SfM reconstruction and learning-based priors over the recently proposed neural radiance fields (NeRF). Unlike existing neural network based optimization method that relies on estimated correspondences, our method directly optimizes over implicit volumes, eliminating the challenging step of matching pixels in indoor scenes. The key to our approach is to utilize the learning-based priors to guide the optimization process of NeRF. Our system firstly adapts a monocular depth network over the target scene by finetuning on its sparse SfM reconstruction. Then, we show that the shape-radiance ambiguity of NeRF still exists in indoor environments and propose to address the issue by employing the adapted depth priors to monitor the sampling process of volume rendering. Finally, a per-pixel confidence map acquired by error computation on the rendered image can be used to further improve the depth quality. Experiments show that our proposed framework significantly outperforms state-of-the-art methods on indoor scenes, with surprising findings presented on the effectiveness of correspondence-based optimization and NeRF-based optimization over the adapted depth priors. In addition, we show that the guided optimization scheme does not sacrifice the original synthesis capability of neural radiance fields, improving the rendering quality on both seen and novel views. Code is available at https://github.com/weiyithu/NerfingMVS.

****

When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes

Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, Gabriel Kreiman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 255-264

Context is of fundamental importance to both human and machine vision; e.g., an object in the air is more likely to be an airplane than a pig. The rich notion of context incorporates several aspects including physics rules, statistical co-occurrences, and relative object sizes, among others. While previous work has focused on crowd-sourced out-of-context photographs from the web to study scene context, controlling the nature and extent of contextual violations has been a daunting task. Here we introduce a diverse, synthetic Out-of-Context Dataset (OCD) with fine-grained control over scene context. By leveraging a 3D simulation engine, we systematically control the gravity, object co-occurrences and relative sizes across 36 object categories in a virtual household environment. We conducted a series of experiments to gain insights into the impact of contextual cues on both human and machine vision using OCD. We conducted psychophysics experiments to establish a human benchmark for out-of-context recognition and then compared it with state-of-the-art computer vision models to quantify the gap between the two. We propose a context-aware recognition transformer model, fusing object and contextual information via multi-head attention. Our model captures useful information for contextual reasoning, enabling human-level performance and better robustness in out-of-context conditions compared to baseline models across OCD and other out-of-context datasets. All source code and data are publicly available at https://github.com/kreimanlab/WhenPigsFlyContext

****

You Don't Only Look Once: Constructing Spatial-Temporal Memory for Integrated 3D Object Detection and Tracking

Jiaming Sun, Yiming Xie, Siyu Zhang, Linghao Chen, Guofeng Zhang, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3185-3194

Humans are able to continuously detect and track surrounding objects by constructing a spatial-temporal memory of the objects when looking around. In contrast, 3D object detectors in existing tracking-by-detection systems often search for objects in every new video frame from scratch, without fully leveraging memory from previous detection results. In this work, we propose a novel system for integrated 3D object detection and tracking, which uses a dynamic object occupancy ma

p and previous object states as spatial-temporal memory to assist object detecti
on in future frames. This memory, together with the ego-motion from back-end odo
metry, guides the detector to achieve more efficient object proposal generation
and more accurate object state estimation. The experiments demonstrate the effec
tiveness of the proposed system and its performance on the ScanNet and KITTI dat
asets. Moreover, the proposed system produces stable bounding boxes and pose tra
jectories over time, while being able to handle occluded and truncated objects.
Code is available at the project page: https://zju3dv.github.io/UDOLO.
****

Learning With Memory-Based Virtual Classes for Deep Metric Learning
Byungsoo Ko, Geonmo Gu, Han-Gyu Kim; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2021, pp. 11792-11801
The core of deep metric learning (DML) involves learning visual similarities in
high-dimensional embedding space. One of the main challenges is to generalize fr
om seen classes of training data to unseen classes of test data. Recent works ha
ve focused on exploiting past embeddings to increase the number of instances for
the seen classes. Such methods achieve performance improvement via augmentation
, while the strong focus on seen classes still remains. This can be undesirable
for DML, where training and test data exhibit entirely different classes. In thi
s work, we present a novel training strategy for DML called MemVir. Unlike previ
ous works, MemVir memorizes both embedding features and class weights to utilize
them as additional virtual classes. The exploitation of virtual classes not onl
y utilizes augmented information for training but also alleviates a strong focus
on seen classes for better generalization. Moreover, we embed the idea of curri
culum learning by slowly adding virtual classes for a gradual increase in learni
ng difficulty, which improves the learning stability as well as the final perfor
mance. MemVir can be easily applied to many existing loss functions without any
modification. Extensive experimental results on famous benchmarks demonstrate th
e superiority of MemVir over state-of-the-art competitors. Code of MemVir is pub
licly available.
****

Excavating the Potential Capacity of Self-Supervised Monocular Depth Estimation
Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, Yangang Cai; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15560-15
569
Self-supervised methods play an increasingly important role in monocular depth e
stimation due to their great potential and low annotation cost. To close the gap
with supervised methods, recent works take advantage of extra constraints, e.g.
, semantic segmentation. However, these methods will inevitably increase the bur
den on the model. In this paper, we show theoretical and empirical evidence that
the potential capacity of self-supervised monocular depth estimation can be exc
avated without increasing this cost. In particular, we propose (1) a novel data
augmentation approach called data grafting, which forces the model to explore mo
re cues to infer depth besides the vertical image position, (2) an exploratory s
elf-distillation loss, which is supervised by the self-distillation label genera
ted by our new post-processing method - selective post-processing, and (3) the f
ull-scale network, designed to endow the encoder with the specialization of dept
h estimation task and enhance the representational power of the model. Extensive
experiments show that our contributions can bring significant performance impro
vement to the baseline with even less computational overhead, and our model, nam
ed EPCDepth, surpasses the previous state-of-the-art methods even those supervis
ed by additional constraints.
****

SPatchGAN: A Statistical Feature Based Discriminator for Unsupervised Image-to-I
mage Translation
Xuning Shao, Weidong Zhang; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2021, pp. 6546-6555
For unsupervised image-to-image translation, we propose a discriminator architec
ture which focuses on the statistical features instead of individual patches. Th
e network is stabilized by distribution matching of key statistical features at

multiple scales. Unlike the existing methods which impose more and more constrai nts on the generator, our method facilitates the shape deformation and enhances the fine details with a greatly simplified framework. We show that the proposed method outperforms the existing state-of-the-art models in various challenging a pplications including selfie-to-anime, male-to-female and glasses removal.
****

Sub-Bit Neural Networks: Learning To Compress and Accelerate Binary Neural Netwo rks
Yikai Wang, Yi Yang, Fuchun Sun, Anbang Yao; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 5360-5369
In the low-bit quantization field, training Binarized Neural Networks (BNNs) is the extreme solution to ease the deployment of deep models on resource-constrain ed devices, having the lowest storage cost and significantly cheaper bit-wise op erations compared to 32-bit floating-point counterparts. In this paper, we intro duce Sub-bit Neural Networks (SNNs), a new type of binary quantization design ta ilored to compress and accelerate BNNs. SNNs are inspired by an empirical observ ation, showing that binary kernels learnt at convolutional layers of a BNN model are likely to be distributed over kernel subsets. As a result, unlike existing methods that binarize weights one by one, SNNs are trained with a kernel-aware o ptimization framework, which exploits binary quantization in the fine-grained co nvolutional kernel space. Specifically, our method includes a random sampling st ep generating layer-specific subsets of the kernel space, and a refinement step learning to adjust these subsets of binary kernels via optimization. Experiments on visual recognition benchmarks and the hardware deployment on FPGA validate t he great potentials of SNNs. For instance, on ImageNet, SNNs of ResNet-18/ResNet -34 with 0.56-bit weights achieve 3.13/3.33 times runtime speed-up and 1.8 times compression over conventional BNNs with moderate drops in recognition accuracy. Promising results are also obtained when applying SNNs to binarize both weights and activations. Our code is available at https://github.com/yikaiw/SNN.
****

Interacting Two-Hand 3D Pose and Shape Reconstruction From Single Color Image
Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, Hon gan Wang; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2021, pp. 11354-11363
In this paper, we propose a novel deep learning framework to reconstruct 3D hand poses and shapes of two interacting hands from a single color image. Previous m ethods designed for single hand cannot be easily applied for the two hand scenar io because of the heavy inter-hand occlusion and larger solution space. In order to address the occlusion and similar appearance between hands that may confuse the network, we design a hand pose-aware attention module to extract features as sociated to each individual hand respectively. We then leverage the two hand con text presented in interaction and propose a context-aware cascaded refinement th at improves the hand pose and shape accuracy of each hand conditioned on the con text between interacting hands. Extensive experiments on the main benchmark data sets demonstrate that our method predicts accurate 3D hand pose and shape from s ingle color image, and achieves the state-of-the-art performance. Code is availa ble in project webpage https://baowenz.github.io/Intershape/.
****

CODEs: Chamfer Out-of-Distribution Examples Against Overconfidence Issue
Keke Tang, Dingruibo Miao, Weilong Peng, Jianpeng Wu, Yawen Shi, Zhaoquan Gu, Zh ihong Tian, Wenping Wang; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2021, pp. 1153-1162
Overconfident predictions on out-of-distribution (OOD) samples is a thorny issue for deep neural networks. The key to resolve the OOD overconfidence issue inher ently is to build a subset of OOD samples and then suppress predictions on them. This paper proposes the Chamfer OOD examples (CODEs), whose distribution is clo se to that of in-distribution samples, and thus could be utilized to alleviate t he OOD overconfidence issue effectively by suppressing predictions on them. To o btain CODEs, we first generate seed OOD examples via slicing&splicing operations on in-distribution samples from different categories, and then feed them to the

Chamfer generative adversarial network for distribution transformation, without accessing to any extra data. Training with suppressing predictions on CODEs is validated to alleviate the OOD overconfidence issue largely without hurting classification accuracy, and outperform the state-of-the-art methods. Besides, we demonstrate CODEs are useful for improving OOD detection and classification.

****

Lifelong Infinite Mixture Model Based on Knowledge-Driven Dirichlet Process
Fei Ye, Adrian G. Bors; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10695-10704

Recent research efforts in lifelong learning propose to grow a mixture of models to adapt to an increasing number of tasks. The proposed methodology shows promising results in overcoming catastrophic forgetting. However, the theory behind these successful models is still not well understood. In this paper, we perform the theoretical analysis for lifelong learning models by deriving the risk bounds based on the discrepancy distance between the probabilistic representation of data generated by the model and that corresponding to the target dataset. Inspired by the theoretical analysis, we introduce a new lifelong learning approach, namely the Lifelong Infinite Mixture (LIMix) model, which can automatically expand its network architectures or choose an appropriate component to adapt its parameters for learning a new task, while preserving its previously learnt information. We propose to incorporate the knowledge by means of Dirichlet processes by using a gating mechanism which computes the dependence between the knowledge learnt previously and stored in each component, and a new set of data. Besides, we train a compact Student model which can accumulate cross-domain representations over time and make quick inferences. The code is available at https://github.com/dtuzi123/Lifelong-infinite-mixture-model.

****

LayoutTransformer: Layout Generation and Completion With Self-Attention
Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S. Davis, Vijay Mahadevan, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1004-1014

We address the problem of scene layout generation for diverse domains such as images, mobile applications, documents, and 3D objects. Most complex scenes, natural or human-designed, can be expressed as a meaningful arrangement of simpler compositional graphical primitives. Generating a new layout or extending an existing layout requires understanding the relationships between these primitives. To do this, we propose LayoutTransformer, a novel framework that leverages self-attention to learn contextual relationships between layout elements and generate novel layouts in a given domain. Our framework allows us to generate a new layout either from an empty set or from an initial seed set of primitives, and can easily scale to support an arbitrary of primitives per layout. Furthermore, our analyses show that the model is able to automatically capture the semantic properties of the primitives. We propose simple improvements in both representation of layout primitives, as well as training methods to demonstrate competitive performance in very diverse data domains such as object bounding boxes in natural images (COCO bounding box), documents (PubLayNet), mobile applications (RICO dataset) as well as 3D shapes (Part-Net). Code and other materials will be made available at https://kampta.github.io/layout.

****

The Power of Points for Modeling Humans in Clothing
Qianli Ma, Jinlong Yang, Siyu Tang, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10974-10984

Currently it requires an artist to create 3D human avatars with realistic clothing that can move naturally. Despite progress on 3D scanning and modeling of human bodies, there is still no technology that can easily turn a static scan into an animatable avatar. Automating the creation of such avatars would enable many applications in games, social networking, animation, and AR/VR to name a few. The key problem is one of representation. Standard 3D meshes are widely used in modeling the minimally-clothed body but do not readily capture the complex topology of clothing. Recent interest has shifted to implicit surface models for this ta

sk but they are computationally heavy and lack compatibility with existing 3D to ols. What is needed is a 3D representation that can capture varied topology at h igh resolution and that can be learned from data. We argue that this representat ion has been with us all along --- the point cloud. Point clouds have properties of both implicit and explicit representations that we exploit to model 3D garme nt geometry on a human body. We train a neural network with a novel local clothi ng geometric feature to represent the shape of different outfits. The network is trained from 3D point clouds of many types of clothing, on many bodies, in many poses, and learns to model pose-dependent clothing deformations. The geometry f eature can be optimized to fit a previously unseen scan of a person in clothing, enabling the scan to be reposed realistically. Our model demonstrates superior quantitative and qualitative results in both multi-outfit modeling and unseen ou tfit animation. The code is available for research purposes at https://qianlim.g ithub.io/POP.

****

Physics-Enhanced Machine Learning for Virtual Fluorescence Microscopy

Colin L. Cooke, Fanjie Kong, Amey Chaware, Kevin C. Zhou, Kanghyun Kim, Rong Xu, D. Michael Ando, Samuel J. Yang, Pavan Chandra Konda, Roarke Horstmeyer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3803-3813

This paper introduces a new method of data-driven microscope design for virtual fluorescence microscopy. We use a deep neural network (DNN) to effectively desig n optical patterns for specimen illumination that substantially improve upon the ability to infer fluorescence image information from unstained microscope image s. To achieve this design, we include an illumination model within the DNN's fir st layers that is jointly optimized during network training. We validated our me thod on two different experimental setups, with different magnifications and sam ple types, to show a consistent improvement in performance as compared to conven tional microscope imaging methods. Additionally, to understand the importance of learned illumination on the inference task, we varied the number of illuminatio n patterns being optimized (and thus the number of unique images captured) and a nalyzed how the structure of the patterns changed as their number increased. Thi s work demonstrates the power of programmable optical elements at enabling bette r machine learning algorithm performance and at providing physical insight into next generation of machine-controlled imaging systems.

****

Dynamic Attentive Graph Learning for Image Restoration

Chong Mou, Jian Zhang, Zhuoyuan Wu; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2021, pp. 4328-4337

Non-local self-similarity in natural images has been verified to be an effective prior for image restoration. However, most existing deep non-local methods assi gn a fixed number of neighbors for each query item, neglecting the dynamics of n on-local correlations. Moreover, the non-local correlations are usually based on pixels, prone to be biased due to image degradation. To rectify these weaknesse s, in this paper, we propose a dynamic attentive graph learning model (DAGL) to explore the dynamic non-local property on patch level for image restoration. Spe cifically, we propose an improved graph model to perform patch-wise graph convol ution with a dynamic and adaptive number of neighbors for each node. In this way , image content can adaptively balance over-smooth and over-sharp artifacts thro ugh the number of its connected neighbors, and the patch-wise non-local correlat ions can enhance the message passing process. Experimental results on various im age restoration tasks: synthetic image denoising, real image denoising, image de mosaicing, and compression artifact reduction show that our DAGL can produce sta te-of-the-art results with superior accuracy and visual quality. The source code is available at https://github.com/jianzhangcs/DAGL.

****

Adversarial Unsupervised Domain Adaptation With Conditional and Label Shift: Inf er, Align and Iterate

Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C.-C. Jay Kuo, George s El Fakhri, Jonghye Woo; Proceedings of the IEEE/CVF International Conference o

n Computer Vision (ICCV), 2021, pp. 10367-10376
In this work, we propose an adversarial unsupervised domain adaptation (UDA) app roach with the inherent conditional and label shifts, in which we aim to align t he distributions w.r.t. both p(x|y) and p(y). Since the label is inaccessible in the target domain, the conventional adversarial UDA assumes p(y) is invariant a cross domains, and relies on aligning p(x) as an alternative to the p(x|y) align ment. To address this, we provide a thorough theoretical and empirical analysis of the conventional adversarial UDA methods under both conditional and label shi fts, and propose a novel and practical alternative optimization scheme for adver sarial UDA. Specifically, we infer the marginal p(y) and align p(x|y) iterativel y in the training, and precisely align the posterior p(y|x) in testing. Our expe rimental results demonstrate its effectiveness on both classification and segmen tation UDA, and partial UDA.
****

Exploring Geometry-Aware Contrast and Clustering Harmonization for Self-Supervis ed 3D Object Detection
Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, Luc Van Gool; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 3293-3302
Current 3D object detection paradigms highly rely on extensive annotation effort s, which makes them not practical in many real-world industrial applications. In spired by that a human driver can keep accumulating experiences from self-explor ing the roads without any tutor's guidance, we first step forwards to explore a simple yet effective self-supervised learning framework tailored for LiDAR-based 3D object detection. Although the self-supervised pipeline has achieved great s uccess in 2D domain, the characteristic challenges (e.g., complex geometry struc ture and various 3D object views) encountered in the 3D domain hinder the direct adoption of existing techniques that often contrast the 2D augmented data or cl uster single-view features. Here we present a novel self-supervised 3D Object de tection framework that seamlessly integrates the geometry-aware contrast and clu stering harmonization to lift the unsupervised 3D representation learning, named GCC-3D. First, GCC-3D introduces a Geometric-Aware Contrastive objective to lea rn spatial-sensitive local structure representation. This objective enforces the spatially-closed voxels to have high feature similarity. Second, a Pseudo-Insta nce Clustering harmonization mechanism is proposed to encourage that different v iews of pseudo-instances should have consistent similarities to clustering proto type centers. This module endows our model semantic discriminative capacity. Ext ensive experiments demonstrate our GCC-3D achieves significant performance impro vement on data-efficient 3D object detection benchmarks (nuScenes and Waymo). Mo reover, our GCC-3D framework can achieve state-of-the art performances on all po pular 3D object detection benchmarks.
****

Learning Meta-Class Memory for Few-Shot Semantic Segmentation
Zhonghua Wu, Xiangxi Shi, Guosheng Lin, Jianfei Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 517-526
Currently, the state-of-the-art methods treat few-shot semantic segmentation tas k as a conditional foreground-background segmentation problem, assuming each cla ss is independent. In this paper, we introduce the concept of meta-class, which is the meta information (e.g. certain middle-level features) shareable among all classes. To explicitly learn meta-class representations in few-shot segmentatio n task, we propose a novel Meta-class Memory based few-shot segmentation method (MM-Net), where we introduce a set of learnable memory embeddings to memorize th e meta-class information during the base class training and transfer to novel cl asses during the inference stage. Moreover, for the k-shot scenario, we propose a novel image quality measurement module to select images from the set of suppor t images. A high-quality class prototype could be obtained with the weighted sum of support image features based on the quality measure. Experiments on both PAS CAL-5^i and COCO datasets show that our proposed method is able to achieve state -of-the-art results in both 1-shot and 5-shot settings. Particularly, our propos ed MM-Net achieves 37.5% mIoU on the COCO dataset in 1-shot setting, which is 5.

1% higher than the previous state-of-the-art.
****
Semi-Supervised Active Learning With Temporal Output Discrepancy
Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, Dejing Dou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3447-3456
While deep learning succeeds in a wide range of tasks, it highly depends on the massive collection of annotated data which is expensive and time-consuming. To lower the cost of data annotation, active learning has been proposed to interactively query an oracle to annotate a small proportion of informative samples in an unlabeled dataset. Inspired by the fact that the samples with higher loss are usually more informative to the model than the samples with lower loss, in this paper we present a novel deep active learning approach that queries the oracle for data annotation when the unlabeled sample is believed to incorporate high loss. The core of our approach is a measurement Temporal Output Discrepancy (TOD) that estimates the sample loss by evaluating the discrepancy of outputs given by models at different optimization steps. Our theoretical investigation shows that TOD lower-bounds the accumulated sample loss thus it can be used to select informative unlabeled samples. On basis of TOD, we further develop an effective unlabeled data sampling strategy as well as an unsupervised learning criterion that enhances model performance by incorporating the unlabeled data. Due to the simplicity of TOD, our active learning approach is efficient, flexible, and task-agnostic. Extensive experimental results demonstrate that our approach achieves superior performances than the state-of-the-art active learning methods on image classification and semantic segmentation tasks.
****
Learning Cross-Modal Contrastive Features for Video Domain Adaptation
Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, Manmohan Chandraker; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13618-13627
Learning transferable and domain adaptive feature representations from videos is important for video-relevant tasks such as action recognition. Existing video domain adaptation methods mainly rely on adversarial feature alignment, which has been derived from the RGB image space. However, video data is usually associated with multi-modal information, e.g., RGB and optical flow, and thus it remains a challenge to design a better method that considers the crossmodal inputs under the cross-domain adaptation setting. To this end, we propose a unified framework for video domain adaptation, which simultaneously regularizes cross-modal and cross-domain feature representations. Specifically, we treat each modality in a domain as a view and leverage the contrastive learning technique with properly designed sampling strategies. As a result, our objectives regularize feature spaces, which originally lack the connection across modalities or have less alignment across domains. We conduct experiments on domain adaptive action recognition benchmark datasets, i.e., UCF, HMDB and EPIC-Kitchens, and demonstrate the effectiveness of our individual components against state-of-the-art algorithms.
****
Energy-Based Open-World Uncertainty Modeling for Confidence Calibration
Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, Dongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9302-9311
Confidence calibration is of great importance to ensure the reliability of decisions made by machine learning systems. However, discriminative classifiers based on deep neural networks are often criticized for producing overconfident predictions that fail to reflect the true correctness likelihood of classification accuracy. We argue that such an inability to model uncertainty is mainly caused by the closed-world nature in softmax: a model trained by the cross-entropy loss will be forced to classify the input into one of K pre-defined categories with high probability. To address this problem, we for the first time propose a novel K+1-way softmax formulation, which incorporates the modeling of open-world uncertainty as to the extra dimension. To unify the learning of the original K-way clas

sification task and the extra dimension that models uncertainty, we (1)propose a novel energy-based objective function, and moreover, (2) theoretically prove that optimizing such an objective essentially forces the extra dimension to capture the marginal data distribution. Extensive experiments show that our approach, Energy-based Open-World Softmax (EOW-Softmax), is superior to existing state-of-the-art methods in improving confidence calibration.
****

Sat2Vid: Street-View Panoramic Video Synthesis From a Single Satellite Image
Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Rongjun Qin, Marc Pollefeys, Martin R. Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12436-12445
We present a novel method for synthesizing both temporally and geometrically consistent street-view panoramic video from a single satellite image and camera trajectory. Existing cross-view synthesis approaches focus on images, while video synthesis in such a case has not yet received enough attention. For geometrical and temporal consistency, our approach explicitly creates a 3D point cloud representation of the scene and maintains dense 3D-2D correspondences across frames that reflect the geometric scene configuration inferred from the satellite view. As for synthesis in the 3D space, we implement a cascaded network architecture with two hourglass modules to generate point-wise coarse and fine features from semantics and per-class latent vectors, followed by projection to frames and an upsampling module to obtain the final realistic video. By leveraging computed correspondences, the produced street-view video frames adhere to the 3D geometric scene structure and maintain temporal consistency. Qualitative and quantitative experiments demonstrate superior results compared to other state-of-the-art synthesis approaches that either lack temporal consistency or realistic appearance. To the best of our knowledge, our work is the first one to synthesize cross-view images to videos.
****

NAS-OoD: Neural Architecture Search for Out-of-Distribution Generalization
Haoyue Bai, Fengwei Zhou, Lanqing Hong, Nanyang Ye, S.-H. Gary Chan, Zhenguo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8320-8329
Recent advances on Out-of-Distribution (OoD) generalization reveal the robustness of deep learning models against distribution shifts. However, existing works focus on OoD algorithms, such as invariant risk minimization, domain generalization, or stable learning, without considering the influence of deep model architectures on OoD generalization, which may lead to sub-optimal performance. Neural Architecture Search (NAS) methods search for architecture based on its performance on the training data, which may result in poor generalization for OoD tasks. In this work, we propose robust Neural Architecture Search for OoD generalization (NAS-OoD), which optimizes the architecture with respect to its performance on generated OoD data by gradient descent. Specifically, a data generator is learned to synthesize OoD data by maximizing losses computed by different neural architectures, while the goal for architecture search is to find the optimal architecture parameters that minimize the synthetic OoD data losses. The data generator and the neural architecture are jointly optimized in an end-to-end manner, and the minimax training process effectively discovers robust architectures that generalize well for different distribution shifts. Extensive experimental results show that NAS-OoD achieves superior performance on various OoD generalization benchmarks with deep models having a much fewer number of parameters. In addition, on a real industry dataset, the proposed NAS-OoD method reduces the error rate by more than 70% compared with the state-of-the-art method, demonstrating the proposed method's practicality for real applications.
****

Hierarchical Aggregation for 3D Instance Segmentation
Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, Xinggang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15467-15476
Instance segmentation on point clouds is a fundamental task in 3D scene percepti

on. In this work, we propose a concise clustering-based framework named HAIS, wh
ich makes full use of spatial relation of points and point sets. Considering clu
stering-based methods may result in over-segmentation or under-segmentation, we
 introduce the hierarchical aggregation to progressively generate instance propos
als, i.e., point aggregation for preliminarily clustering points to sets and set
 aggregation for generating complete instances from sets. Once the complete 3D i
nstances are obtained, a sub-network of intra-instance prediction is adopted for
 noisy points filtering and mask quality scoring. HAIS is fast (only 410ms per f
rame on Titan X) and does not require non-maximum suppression. It ranks 1st on t
he ScanNet v2 benchmark, achieving the highest 69.9% AP50 and surpassing previou
s state-of-the-art (SOTA) methods by a large margin. Besides, the SOTA results o
n the S3DIS dataset validate the good generalization ability. Code is available
at https://github.com/hustvl/HAIS.
****

Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Stu
dies on Medical Image Classification

Zhuoning Yuan, Yan Yan, Milan Sonka, Tianbao Yang; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 3040-3049

Deep AUC Maximization (DAM) is a new paradigm for learning a deep neural network
 by maximizing the AUC score of the model on a dataset. Most previous works of A
UC maximization focus on the perspective of optimization by designing efficient
stochastic algorithms, and studies on generalization performance of large-scale
DAM on difficult tasks are missing. In this work, we aim to make DAM more practi
cal for interesting real-world applications (e.g., medical image classification)
. First, we propose a new margin-based min-max surrogate loss function for the A
UC score (named as the AUC min-max-margin loss or simply AUC margin loss for sho
rt). It is more robust than the commonly used AUC square loss, while enjoying th
e same advantage in terms of large-scale stochastic optimization. Second, we con
duct extensive empirical studies of our DAM method on four difficult medical ima
ge classification tasks, namely (i) classification of chest x-ray images for ide
ntifying many threatening diseases, (ii) classification of images of skin lesion
s for identifying melanoma, (iii) classification of mammogram for breast cancer
screening, and (iv) classification of microscopic images for identifying tumor t
issue. Our studies demonstrate that the proposed DAM method improves the perform
ance of optimizing cross-entropy loss by a large margin, and also achieves bette
r performance than optimizing the existing AUC square loss on these medical imag
e classification tasks. Specifically, our DAM method has achieved the 1st place
on Stanford CheXpert competition on Aug. 31, 2020. To the best of our knowledge,
 this is the first work that makes DAM succeed on large-scale medical image data
sets. We also conduct extensive ablation studies to demonstrate the advantages o
f the new AUC margin loss over the AUC square loss on benchmark datasets. The pr
oposed method is implemented in our open-sourced library LibAUC (www.libauc.org)
 whose github address is https://github.com/Optimization-AI/LibAUC.
****

A Simple Baseline for Semi-Supervised Semantic Segmentation With Strong Data Aug
mentation

Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, Hao Li; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8229-8238

Recently, significant progress has been made on semantic segmentation. However,
the success of supervised semantic segmentation typically relies on a large amou
nt of labeled data, which is time-consuming and costly to obtain. Inspired by th
e success of semi-supervised learning methods in image classification, here we p
ropose a simple yet effective semi-supervised learning framework for semantic se
gmentation. We demonstrate that the devil is in the details: a set of simple des
ign and training techniques can collectively improve the performance of semi-sup
ervised semantic segmentation significantly. Previous works fail to employ stron
g augmentation in pseudo label learning efficiently, as the large distribution c
hange caused by strong augmentation harms the batch normalization statistics. We
 design a new batch normalization, namely distribution-specific batch normalizat
ion (DSBN) to address this problem and demonstrate the importance of strong augm

entation for semantic segmentation. Moreover, we design a self-correction loss w hich is effective in noise resistance. We conduct a series of ablation studies t o show the effectiveness of each component. Our method achieves state-of-the-art results in the semi-supervised settings on the Cityscapes and Pascal VOC datase ts.
****

Ground-Truth or DAER: Selective Re-Query of Secondary Information
Stephan J. Lemmer, Jason J. Corso; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 703-714
Many vision tasks use secondary information at inference time---a seed---to assi st a computer vision model in solving a problem. For example, an initial boundin g box is needed to initialize visual object tracking. To date, all such work mak es the assumption that the seed is a good one. However, in practice, from crowds ourcing to noisy automated seeds, this is often not the case. We hence propose t he problem of seed rejection---determining whether to reject a seed based on the expected performance degradation when it is provided in place of a gold-standar d seed. We provide a formal definition to this problem, and focus on two meaning ful subgoals: understanding causes of error and understanding the model's respon se to noisy seeds conditioned on the primary input. With these goals in mind, we propose a novel training method and evaluation metrics for the seed rejection p roblem. We then use seeded versions of the viewpoint estimation and fine-grained classification tasks to evaluate these contributions. In these experiments, we show our method can reduce the number of seeds that need to be reviewed for a ta rget performance by over 23% compared to strong baselines.
****

Evidential Deep Learning for Open Set Action Recognition
Wentao Bao, Qi Yu, Yu Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13349-13358
In a real-world scenario, human actions are typically out of the distribution fr om training data, which requires a model to both recognize the known actions and reject the unknown. Different from image data, video actions are more challengi ng to be recognized in an open-set setting due to the uncertain temporal dynamic s and static bias of human actions. In this paper, we propose a Deep Evidential Action Recognition (DEAR) method to recognize actions in an open testing set. Sp ecifically, we formulate the action recognition problem from the evidential deep learning (EDL) perspective and propose a novel model calibration method to regu larize the EDL training. Besides, to mitigate the static bias of video represent ation, we propose a plug-and-play module to debias the learned representation th rough contrastive learning. Experimental results show that our DEAR method achie ves consistent performance gain on multiple mainstream action recognition models and benchmarks. Code and pre-trained models are available at https://www.rit.ed u/actionlab/dear.
****

Perception-Aware Multi-Sensor Fusion for 3D LiDAR Semantic Segmentation
Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, Mingkui Tan; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021 , pp. 16280-16290
3D LiDAR (light detection and ranging) semantic segmentation is important in sce ne understanding for many applications, such as auto-driving and robotics. For e xample, for autonomous cars equipped with RGB cameras and LiDAR, it is crucial t o fuse complementary information from different sensors for robust and accurate segmentation. Existing fusion-based methods, however, may not achieve promising performance due to the vast difference between the two modalities. In this work, we investigate a collaborative fusion scheme called perception-aware multi-sens or fusion (PMF) to exploit perceptual information from two modalities, namely, a ppearance information from RGB images and spatio-depth information from point cl ouds. To this end, we first project point clouds to the camera coordinates to pr ovide spatio-depth information for RGB images. Then, we propose a two-stream net work to extract features from the two modalities, separately, and fuse the featu res by effective residual-based fusion modules. Moreover, we propose additional

perception-aware losses to measure the perceptual difference between the two modalities. Extensive experiments on two benchmark data sets show the superiority of our method. For example, on nuScenes, our PMF outperforms the state-of-the-art method by 0.8% in mIoU.
****

UVStyle-Net: Unsupervised Few-Shot Learning of 3D Style Similarity Measure for B-Reps
Peter Meltzer, Hooman Shayani, Amir Khasahmadi, Pradeep Kumar Jayaraman, Aditya Sanghi, Joseph Lambourne; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9690-9699
Boundary Representations (B-Reps) are the industry standard in 3D Computer Aided Design/Manufacturing (CAD/CAM) and industrial design due to their fidelity in representing stylistic details. However, they have been ignored in the 3D style research. Existing 3D style metrics typically operate on meshes or point clouds, and fail to account for end-user subjectivity by adopting fixed definitions of style, either through crowd-sourcing for style labels or hand-crafted features. We propose UVStyle-Net, a style similarity measure for B-Reps that leverages the style signals in the second order statistics of the activations in a pre-trained (unsupervised) 3D encoder, and learns their relative importance to a subjective end-user through few-shot learning. Our approach differs from all existing data-driven 3D style methods since it may be used in completely unsupervised settings, which is desirable given the lack of publicly available labeled B-Rep datasets. More importantly, the few-shot learning accounts for the inherent subjectivity associated with style. We show quantitatively that our proposed method with B-Reps is able to capture stronger style signals than alternative methods on meshes and point clouds despite its significantly greater computational efficiency. We also show it is able to generate meaningful style gradients with respect to the input shape, and that few-shot learning with as few as two positive examples selected by an end-user is sufficient to significantly improve the style measure. Finally, we demonstrate its efficacy on a large unlabeled public dataset of CAD models. Source code and data are available at https://github.com/AutodeskAILab/UVStyle-Net.
****

End-to-End Dense Video Captioning With Parallel Decoding
Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6847-6857
Dense video captioning aims to generate multiple associated captions with their temporal locations from the video. Previous methods follow a sophisticated "localize-then-describe" scheme, which heavily relies on numerous hand-crafted components. In this paper, we proposed a simple yet effective framework for end-to-end dense video captioning with parallel decoding (PDVC), by formulating the dense caption generation as a set prediction task. In practice, through stacking a newly proposed event counter on the top of a transformer decoder, the PDVC precisely segments the video into a number of event pieces under the holistic understanding of the video content, which effectively increases the coherence and readability of predicted captions. Compared with prior arts, the PDVC has several appealing advantages: (1) Without relying on heuristic non-maximum suppression or a recurrent event sequence selection network to remove redundancy, PDVC directly produces an event set with an appropriate size; (2) In contrast to adopting the two-stage scheme, we feed the enhanced representations of event queries into the localization head and caption head in parallel, making these two sub-tasks deeply interrelated and mutually promoted through the optimization; (3) Without bells and whistles, extensive experiments on ActivityNet Captions and YouCook2 show that PDVC is capable of producing high-quality captioning results, surpassing the state-of-the-art two-stage methods when its localization accuracy is on par with them. Code is available at https://github.com/ttengwang/PDVC.
****

StarEnhancer: Learning Real-Time and Style-Aware Image Enhancement
Yuda Song, Hui Qian, Xin Du; Proceedings of the IEEE/CVF International Conferenc

e on Computer Vision (ICCV), 2021, pp. 4126-4135
Image enhancement is a subjective process whose targets vary with user preferenc es. In this paper, we propose a deep learning-based image enhancement method cov ering multiple tonal styles using only a single model dubbed StarEnhancer. It ca n transform an image from one tonal style to another, even if that style is unse en. With a simple one-time setting, users can customize the model to make the en hanced images more in line with their aesthetics. To make the method more practi cal, we propose a well-designed enhancer that can process a 4K-resolution image over 200 FPS but surpasses the contemporaneous single style image enhancement me thods in terms of PSNR, SSIM, and LPIPS. Finally, our proposed enhancement metho d has good interactability, which allows the user to fine-tune the enhanced imag e using intuitive options.
****

Can Shape Structure Features Improve Model Robustness Under Diverse Adversarial Settings?
Mingjie Sun, Zichao Li, Chaowei Xiao, Haonan Qiu, Bhavya Kailkhura, Mingyan Liu, Bo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7526-7535
Recent studies show that convolutional neural networks (CNNs) are vulnerable und er various settings, including adversarial attacks, common corruptions, and back door attacks. Motivated by the findings that human visual system pays more atten tion to global structure (e.g., shapes) for recognition while CNNs are biased to wards local texture features in images, in this work we aim to analyze whether " edge features" could improve the recognition robustness in these scenarios, and if so, to what extent? To answer these questions and systematically evaluate the global structure features, we focus on shape features and pro-pose two edge-ena bled pipelines EdgeNetRob and Edge-GANRob, forcing the CNNs to rely more on edge features. Specifically, EdgeNetRob and EdgeGANRob first explicitly extract shap e structure features from a given image via an edge detection algorithm. Then Ed geNetRob trains down-stream learning tasks directly on the extracted edge featur es, while EdgeGANRob reconstructs a new image by re-filling the texture informat ion with a trained generative adversarial network (GANs). To reduce the sensitiv ity of edge detection algorithms to perturbations, we additionally propose a rob ust edge detection approach Robust Canny based on vanilla Canny. Based on our ev aluation, we find that EdgeNetRob can help boost model robustness under differ-e nt attack scenarios at the cost of the clean model accuracy. EdgeGANRob, on the other hand, is able to improve the clean model accuracy compared to EdgeNetRob w hile preserving robustness. This shows that given such edge features, how to lev erage them matters for robustness, and it also depends on data properties. Our s ystematic studies on edge structure features under different settings will shed light on future robust feature exploration and optimization.
****

Multi-Class Cell Detection Using Spatial Context Representation
Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee, Raj arsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, Chao Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4005 -4014
In digital pathology, both detection and classification of cells are important f or automatic diagnostic and prognostic tasks. Classifying cells into subtypes, s uch as tumor cells, lymphocytes or stromal cells is particularly challenging. Ex isting methods focus on morphological appearance of individual cells, whereas in practice pathologists often infer cell classes through their spatial context. I n this paper, we propose a novel method for both detection and classification th at explicitly incorporates spatial contextual information. We use the spatial st atistical function to describe local density in both a multi-class and a multi-s cale manner. Through representation learning and deep clustering techniques, we learn advanced cell representation with both appearance and spatial context. On various benchmarks, our method achieves better performance than state-of-the-art s, especially on the classification task.
****

Learning by Aligning: Visible-Infrared Person Re-Identification Using Cross-Modal Correspondences
Hyunjong Park, Sanghoon Lee, Junghyup Lee, Bumsub Ham; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12046-12055
We address the problem of visible-infrared person re-identification (VI-reID), that is, retrieving a set of person images, captured by visible or infrared cameras, in a cross-modal setting. Two main challenges in VI-reID are intra-class variations across person images, and cross-modal discrepancies between visible and infrared images. Assuming that the person images are roughly aligned, previous approaches attempt to learn coarse image- or rigid part-level person representations that are discriminative and generalizable across different modalities. However, the person images, typically cropped by off-the-shelf object detectors, are not necessarily well-aligned, which distract discriminative person representation learning. In this paper, we introduce a novel feature learning framework that addresses these problems in a unified way. To this end, we propose to exploit dense correspondences between cross-modal person images. This allows to address the cross-modal discrepancies in a pixel-level, suppressing modality-related features from person representations more effectively. This also encourages pixel-wise associations between cross-modal local features, further facilitating discriminative feature learning for VI-reID. Extensive experiments and analyses on standard VI-reID benchmarks demonstrate the effectiveness of our approach, which significantly outperforms the state of the art.
****

Localize to Binauralize: Audio Spatialization From Visual Sound Source Localization
Kranthi Kumar Rachavarapu, Aakanksha, Vignesh Sundaresha, A. N. Rajagopalan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1930-1939
Videos with binaural audios provide an immersive viewing experience by enabling 3D sound sensation. Recent works attempt to generate binaural audio in a multimodal learning framework using large quantities of videos with accompanying binaural audio. In contrast, we attempt a more challenging problem -- synthesizing binaural audios for a video with monaural audio in a weakly supervised setting and weakly semi-supervised setting. Our key idea is that any down-stream task that can be solved only using binaural audios can be used to provide proxy supervision for binaural audio generation, thereby reducing the reliance on explicit supervision. In this work, as a proxy-task for weak supervision, we use Sound Source Localization with only audio. We design a two-stage architecture called Localize-to-Binauralize Network (L2BNet). The first stage of L2BNet is a Stereo Generation (SG) network employed to generate two-stream audio from monaural audio using visual frame information as guidance. In the second stage, an Audio Localization (AL) network is designed to use the synthesized two-stream audio to localize sound sources in visual frames. The entire network is trained end-to-end so that the AL network provides necessary supervision for the SG network. We experimentally show that our weakly-supervised framework generates two-stream audio containing binaural cues. Through user study, we further validate that our proposed approach generates binaural-quality audio using as little as 10% of explicit binaural supervision data for the SG network.
****

ALL Snow Removed: Single Image Desnowing Algorithm Using Hierarchical Dual-Tree Complex Wavelet Representation and Contradict Channel Loss
Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I-Hsiang Chen, Jian-Jiun Ding, Sy-Yen Kuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4196-4205
Snow is a highly complicated atmospheric phenomenon that usually contains snowflake, snow streak, and veiling effect (similar to the haze or the mist). In this literature, we propose a single image desnowing algorithm to address the diversity of snow particles in shape and size. First, to better represent the complex snow shape, we apply the dual-tree wavelet transform and propose a complex wavelet loss in the network. Second, we propose a hierarchical decomposition paradigm

in our network for better understanding the different sizes of snow particles. L
ast, we propose a novel feature called the contradict channel (CC) for the snow
scenes. We find that the regions containing the snow particles tend to have high
er intensity in the CC than that in the snow-free regions. We leverage this disc
riminative feature to construct the contradict channel loss for improving the pe
rformance of snow removal. Moreover, due to the limitation of existing snow data
sets, to simulate the snow scenarios comprehensively, we propose a large-scale d
ataset called Comprehensive Snow Dataset (CSD). Experimental results show that t
he proposed method can favorably outperform existing methods in three synthetic
datasets and real-world datasets. The code and dataset are released in https://g
ithub.com/weitingchen83/ICCV2021-Single-Image-Desnowing-HDCWNet.
****

MINE: Towards Continuous Depth MPI With NeRF for Novel View Synthesis
Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, Gim Hee Lee; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 12578-12588
In this paper, we propose MINE to perform novel view synthesis and depth estimat
ion via dense 3D reconstruction from a single image. Our approach is a continuou
s depth generalization of the Multiplane Images (MPI) by introducing the NEural
radiance fields (NeRF). Given a single image as input, MINE predicts a 4-channel
 image (RGB and volume density) at arbitrary depth values to jointly reconstruct
 the camera frustum and fill in occluded contents. The reconstructed and inpaint
ed frustum can then be easily rendered into novel RGB or depth views using diffe
rentiable rendering. Extensive experiments on RealEstate10K, KITTI and Flowers L
ight Fields show that our MINE outperforms state-of-the-art by a large margin in
 novel view synthesis. We also achieve competitive results in depth estimation o
n iBims-1 and NYU-v2 without annotated depth supervision. Our source code is ava
ilable at https://github.com/vincentfung13/MINE
****

LoFGAN: Fusing Local Representations for Few-Shot Image Generation
Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, Yang Gao; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2021, pp. 8463-8471
Given only a few available images for a novel unseen category, few-shot image ge
neration aims to generate more data for this category. Previous works attempt to
 globally fuse these images by using adjustable weighted coefficients. However,
there is a serious semantic misalignment between different images from a global
perspective, making these works suffer from poor generation quality and diversit
y. To tackle this problem, we propose a novel Local-Fusion Generative Adversaria
l Network (LoFGAN) for few-shot image generation. Instead of using these availab
le images as a whole, we first randomly divide them into a base image and severa
l reference images. Next, LoFGAN matches local representations between the base
and reference images based on semantic similarities and replaces the local featu
res with the closest related local features. In this way, LoFGAN can produce mor
e realistic and diverse images at a more fine-grained level, and simultaneously
enjoy the characteristic of semantic alignment. Furthermore, a local reconstruct
ion loss is also proposed, which can provide better training stability and gener
ation quality. We conduct extensive experiments on three datasets, which success
fully demonstrates the effectiveness of our proposed method for few-shot image g
eneration and downstream visual applications with limited data. Code is availabl
e at https://github.com/edward3862/LoFGAN-pytorch.
****

Grafit: Learning Fine-Grained Image Representations With Coarse Labels
Hugo Touvron, Alexandre Sablayrolles, Matthijs Douze, Matthieu Cord, Hervé Jégou
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2021, pp. 874-884
This paper tackles the problem of learning a finer representation than the one p
rovided by training labels. This enables fine-grained category retrieval of imag
es in a collection annotated with coarse labels only. Our network is learned wit
h a nearest-neighbor classifier objective, and an instance loss inspired by self
-supervised learning. By jointly leveraging the coarse labels and the underlying

fine-grained latent space, it significantly improves the accuracy of category-level retrieval methods. Our strategy outperforms all competing methods for retrieving or classifying images at a finer granularity than that available at train time. It also improves the accuracy for transfer learning tasks to fine-grained datasets.
****

Rethinking the Truly Unsupervised Image-to-Image Translation
Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, Hyunjung Shim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14154-14163
Every recent image-to-image translation model inherently requires either image-level (i.e. input-output pairs) or set-level (i.e. domain labels) supervision. However, even set-level supervision can be a severe bottleneck for data collection in practice. In this paper, we tackle image-to-image translation in a fully unsupervised setting, i.e., neither paired images nor domain labels. To this end, we propose a truly unsupervised image-to-image translation model (TUNIT) that simultaneously learns to separate image domains and translates input images into the estimated domains. Experimental results show that our model achieves comparable or even better performance than the set-level supervised model trained with full labels, generalizes well on various datasets, and is robust against the choice of hyperparameters (e.g. the preset number of pseudo domains). Furthermore, TUNIT can be easily extended to semi-supervised learning with a few labeled data.
****

Point-Based Modeling of Human Clothing
Ilya Zakharkin, Kirill Mazur, Artur Grigorev, Victor Lempitsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14718-14727
We propose a new approach to human clothing modeling based on point clouds. Within this approach, we learn a deep model that can predict point clouds of various outfits, for various human poses, and for various human body shapes. Notably, outfits of various types and topologies can be handled by the same model. Using the learned model, we can infer the geometry of new outfits from as little as a single image, and perform outfit retargeting to new bodies in new poses. We complement our geometric model with appearance modeling that uses the point cloud geometry as a geometric scaffolding and employs neural point-based graphics to capture outfit appearance from videos and to re-render the captured outfits. We validate both geometric modeling and appearance modeling aspects of the proposed approach against recently proposed methods and establish the viability of point-based clothing modeling.
****

Equivariant Imaging: Learning Beyond the Range Space
Dongdong Chen, Julián Tachella, Mike E. Davies; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4379-4388
In various imaging problems, we only have access to compressed measurements of the underlying signals, hindering most learning-based strategies which usually require pairs of signals and associated measurements for training. Learning only from compressed measurements is impossible in general, as the compressed observations do not contain information outside the range of the forward sensing operator. We propose a new end-to-end self-supervised framework that overcomes this limitation by exploiting the equivariances present in natural signals. Our proposed learning strategy performs as well as fully supervised methods. Experiments demonstrate the potential of this framework on inverse problems including sparse-view X-ray computed tomography on real clinical data and image inpainting on natural images. Code has been made available at: https://github.com/edongdongchen/EI.
****

Mitigating Intensity Bias in Shadow Detection via Feature Decomposition and Reweighting
Lei Zhu, Ke Xu, Zhanghan Ke, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4702-4711
While CNNs achieved remarkable progress in shadow detection, they tend to make m

istakes in dark non-shadow regions and relatively bright shadow regions. They ar e also susceptible to brightness change. These two phenomenons reveal that deep shadow detectors heavily depend on the intensity cue, which we refer to as inten sity bias. In this paper, we propose a novel feature decomposition and reweighti ng scheme to mitigate this intensity bias, in which multi-level integrated featu res are decomposed into intensity-variant and intensity-invariant components thr ough self-supervision. By reweighting these two types of features, our method ca n reallocate the attention to the corresponding latent semantics and achieves ba lanced exploitation of them. Extensive experiments on three popular datasets sho w that the proposed method outperforms state-of-the-art shadow detectors.
****

Joint Representation Learning and Novel Category Discovery on Single- and Multi-Modal Data
Xuhui Jia, Kai Han, Yukun Zhu, Bradley Green; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2021, pp. 610-619
This paper studies the problem of novel category discovery on single- and multi-modal data with labels from different but relevant categories. We present a gene ric, end-to-end framework to jointly learn a reliable representation and assign clusters to unlabelled data. To avoid over-fitting the learnt embedding to label led data, we take inspiration from self-supervised representation learning by no ise-contrastive estimation and extend it to jointly handle labelled and unlabell ed data. In particular, we propose using category discrimination on labelled dat a and cross-modal discrimination on multi-modal data to augment instance discrim ination used in conventional contrastive learning approaches. We further employ Winner-Take-All (WTA) hashing algorithm on the shared representation space to ge nerate pairwise pseudo labels for unlabelled data to better predict cluster assi gnments. We thoroughly evaluate our framework on large-scale multi-modal video b enchmarks Kinetics-400 and VGG-Sound, and image benchmarks CIFAR10, CIFAR100 and ImageNet, obtaining state-of-the-art results.
****

Sparse Needlets for Lighting Estimation With Spherical Transport Loss
Fangneng Zhan, Changgong Zhang, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Visi on (ICCV), 2021, pp. 12830-12839
Accurate lighting estimation is challenging yet critical to many computer vision and computer graphics tasks such as high-dynamic-range (HDR) relighting. Existi ng approaches model lighting in either frequency domain or spatial domain which is insufficient to represent the complex lighting conditions in scenes and tends to produce inaccurate estimation. This paper presents NeedleLight, a new lighti ng estimation model that represents illumination with needlets and allows lighti ng estimation in both frequency domain and spatial domain jointly. An optimal th resholding function is designed to achieve sparse needlets which trims redundant lighting parameters and demonstrates superior localization properties for illum ination representation. In addition, a novel spherical transport loss is designe d based on optimal transport theory which guides to regress lighting representat ion parameters with consideration of the spatial information. Furthermore, we pr opose a new metric that is concise yet effective by directly evaluating the esti mated illumination maps rather than rendered images. Extensive experiments show that NeedleLight achieves superior lighting estimation consistently across multi ple evaluation metrics as compared with state-of-the-art methods.
****

CANet: A Context-Aware Network for Shadow Removal
Zipei Chen, Chengjiang Long, Ling Zhang, Chunxia Xiao; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2021, pp. 4743-4752
In this paper, we propose a novel two-stage context-aware network named CANet fo r shadow removal, in which the contextual information from non-shadow regions is transferred to shadow regions at the embedded feature spaces. At Stage-I, we pr opose a contextual patch matching module to generate a set of potential matching pairs of shadow and non-shadow patches. Combined with the potential contextual relationships between shadow and non-shadow regions, our well-designed contextua

l feature transfer (CFT) mechanism can transfer contextual information from non-shadow to shadow regions at different scales. With the reconstructed feature maps, we remove shadows at L and A/B channels separately. At Stage-II, we use an encoder-decoder to refine current results and generate the final shadow removal results. We evaluate our proposed CANet on two benchmark datasets and some real-world shadow images with complex scenes. Extensive experiment results strongly demonstrate the efficacy of our proposed CANet and exhibit superior performance to state-of-the-arts.

****

Semantic Perturbations With Normalizing Flows for Improved Generalization

Oguz Kaan Yüksel, Sebastian U. Stich, Martin Jaggi, Tatjana Chavdarova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6619-6629

Data augmentation is a widely adopted technique for avoiding overfitting when training deep neural networks. However, this approach requires domain-specific knowledge and is often limited to a fixed set of hard-coded transformations. Recently, several works proposed to use generative models for generating semantically meaningful perturbations to train a classifier. However, because accurate encoding and decoding is critical, these methods, which use architectures that approximate the latent-variable inference, remained limited to pilot studies on small datasets. Exploiting the exactly reversible encoder-decoder structure of normalizing flows, we perform on-manifold perturbations in the latent space to define fully unsupervised data augmentations. We demonstrate that such perturbations match the performance of advanced data augmentation techniques---reaching 96.6% test accuracy for CIFAR-10 using ResNet-18 and outperform existing methods, particularly in low data regimes---yielding 10--25% relative improvement of test accuracy from classical training. We find that our latent adversarial perturbations adaptive to the classifier throughout its training are most effective, yielding the first test accuracy improvement results on real-world datasets---CIFAR-10/100---via latent-space perturbations.

****

Audio-Visual Floorplan Reconstruction

Senthil Purushwalkam, Sebastià Vicenc Amengual Garí, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, Kristen Grauman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1183-1192

Given only a few glimpses of an environment, how much can we infer about its entire floorplan? Existing methods can map only what is visible or immediately apparent from context, and thus require substantial movements through a space to fully map it. We explore how both audio and visual sensing together can provide rapid floorplan reconstruction from limited viewpoints. Audio not only helps sense geometry outside the camera's field of view, but it also reveals the existence of distant freespace (e.g., a dog barking in another room) and suggests the presence of rooms not visible to the camera (e.g., a dishwasher humming in what must be the kitchen to the left). We introduce AV-Map, a novel multi-modal encoder-decoder framework that reasons jointly about audio and vision to reconstruct a floorplan from a short input video sequence. We train our model to predict both the interior structure of the environment and the associated rooms' semantic labels. Our results on 85 large real-world environments show the impact: with just a few glimpses spanning 26% of an area, we can estimate the whole area with 66% accuracy---substantially better than the state of the art approach for extrapolating visual maps.

****

MonoIndoor: Towards Good Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments

Pan Ji, Runze Li, Bir Bhanu, Yi Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12787-12796

Self-supervised depth estimation for indoor environments is more challenging than its outdoor counterpart in at least the following two aspects: (i) the depth range of indoor sequences varies a lot across different frames, making it difficult for the depth network to induce consistent depth cues, whereas the maximum di

stance in outdoor scenes mostly stays the same as the camera usually sees the sk
y; (ii) the indoor sequences contain much more rotational motions, which cause d
ifficulties for the pose network, while the motions of outdoor sequences are pre
-dominantly translational, especially for driving datasets such as KITTI. In thi
s paper, special considerations are given to those challenges and a set of good
practices are consolidated for improving the performance of self-supervised mono
cular depth estimation in indoor environments. The proposed method mainly consis
ts of two novel modules, i.e., a depth factorization module and a residual pose
estimation module, each of which is designed to respectively tackle the aforemen
tioned challenges. The effectiveness of each module is shown through a carefully
 conducted ablation study and the demonstration of the state-of-the-art performa
nce on three indoor datasets, i.e., EuRoC, NYUv2 and 7-Scenes.
****

Common Objects in 3D: Large-Scale Learning and Evaluation of Real-Life 3D Catego
ry Reconstruction

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick La
batut, David Novotny; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2021, pp. 10901-10911

Traditional approaches for learning 3D object categories have been predominantly
 trained and evaluated on synthetic datasets due to the unavailability of real 3
D-annotated category-centric data. Our main goal is to facilitate advances in th
is field by collecting real-world data in a magnitude similar to the existing sy
nthetic counterparts. The principal contribution of this work is thus a large-sc
ale dataset, called Common Objects in 3D, with real multi-view images of object
categories annotated with camera poses and ground truth 3D point clouds. The dat
aset contains a total of 1.5 million frames from nearly 19,000 videos capturing
objects from 50 MS-COCO categories and, as such, it is significantly larger than
 alternatives both in terms of the number of categories and objects. We exploit
this new dataset to conduct one of the first large-scale "in-the-wild" evaluatio
ns of several new-view-synthesis and category-centric 3D reconstruction methods.
 Finally, we contribute NerFormer - a novel neural rendering method that leverag
es the powerful Transformer to reconstruct an object given a small number of its
 views.
****

Reconstructing Hand-Object Interactions in the Wild

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, Jitendra Malik; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12417-124
26

We study the problem of understanding hand-object interactions from 2D images in
 the wild. This requires reconstructing both the hand and the object in 3D, whic
h is challenging because of the mutual occlusion between the hand and the object
. In this paper we make two main contributions: (1) a novel reconstruction techn
ique, RHO (Reconstructing Hands and Objects), which reconstructs 3D models of bo
th the hand and the object leveraging the 2D image cues and 3D contact priors; (
2) a dataset MOW (Manipulating Objects in the Wild) of 500 examples of hand-obje
ct interaction images that have been "3Dfied" with the help of the RHO technique
. Overall our dataset contains 121 distinct object categories, with a much great
er diversity of manipulation actions, than in previous datasets.
****

TOOD: Task-Aligned One-Stage Object Detection

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, Weilin Huang; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3
510-3519

One-stage object detection is commonly implemented by optimizing two sub-tasks:
object classification and localization, using heads with two parallel branches,
which might lead to a certain level of spatial misalignment in predictions betwe
en the two tasks. In this work, we propose a Task-aligned One-stage Object Detec
tion (TOOD) that explicitly aligns the two tasks in a learning-based manner. Fir
st, we design a novel Task-aligned Head (T-Head) which offers a better balance b
etween learning task-interactive and task-specific features, as well as a greate

r flexibility to learn the alignment via a task-aligned predictor. Second, we pr
opose Task Alignment Learning (TAL) to explicitly pull closer (or even unify) th
e optimal anchors for the two tasks during training via a designed sample assign
ment scheme and a task-aligned loss. Extensive experiments are conducted on MS-C
OCO, where TOOD achieves a 51.1 AP at single-model single-scale testing. This su
rpasses the recent one-stage detectors by a large margin, such as ATSS (47.7 AP)
, GFL (48.2 AP), and PAA (49.0 AP), with fewer parameters and FLOPs. Qualitative
 results also demonstrate the effectiveness of TOOD for better aligning the task
s of object classification and localization. Code is available at https://github
.com/fcjian/TOOD.
****

Generalizable Mixed-Precision Quantization via Attribution Rank Preservation
Ziwei Wang, Han Xiao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2021, pp. 5291-5300
In this paper, we propose a generalizable mixed-precision quantization (GMPQ) me
thod for efficient inference. Conventional methods require the consistency of da
tasets for bitwidth search and model deployment to guarantee the policy optimali
ty, leading to heavy search cost on challenging largescale datasets in realistic
 applications. On the contrary, our GMPQ searches the mixed-quantization policy
that can be generalized to largescale datasets with only a small amount of data,
 so that the search cost is significantly reduced without performance degradatio
n. Specifically, we observe that locating network attribution correctly is gener
al ability for accurate visual analysis across different data distribution. Ther
efore, despite of pursuing higher model accuracy and complexity, we preserve att
ribution rank consistency between the quantized models and their full-precision
counterparts via efficient capacity-aware attribution imitation for generalizabl
e mixed-precision quantization strategy search. Extensive experiments show that
our method obtains competitive accuracy-complexity trade-off compared with the s
tate-of-the-art mixed-precision networks in significantly reduced search cost. T
he code is available at https://github.com/ZiweiWangTHU/GMPQ.git.
****

LabOR: Labeling Only if Required for Domain Adaptive Semantic Segmentation
Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, Kwanyong Park, In So Kweon;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2021, pp. 8588-8598
Unsupervised Domain Adaptation (UDA) for semantic segmentation has been actively
 studied to mitigate the domain gap between label-rich source data and unlabeled
 target data. Despite these efforts, UDA still has a long way to go to reach the
 fully supervised performance. To this end, we propose a Labeling Only if Requir
ed strategy, LabOR, where we introduce a human-in-the-loop approach to adaptivel
y give scarce labels to points that a UDA model is uncertain about. In order to
find the uncertain points, we generate an inconsistency mask using the proposed
adaptive pixel selector and we label these segment-based regions to achieve near
 supervised performance with only a small fraction (about 2.2%) ground truth poi
nts, which we call "Segment based Pixel-Labeling (SPL)." To further reduce the e
fforts of the human annotator, we also propose "Point based Pixel-Labeling (PPL)
," which finds the most representative points for labeling within the generated
inconsistency mask. This reduces efforts from 2.2% segment label to 40 points la
bel while minimizing performance degradation. Through extensive experimentation,
 we show the advantages of this new framework for domain adaptive semantic segme
ntation while minimizing human labor costs.
****

SPEC: Seeing People in the Wild With an Estimated Camera
Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges,
Michael J. Black; Proceedings of the IEEE/CVF International Conference on Comput
er Vision (ICCV), 2021, pp. 11035-11045
Due to the lack of camera parameter information for in-the-wild images, existing
 3D human pose and shape (HPS) estimation methods make several simplifying assum
ptions: weak-perspective projection, large constant focal length, and zero camer
a rotation. These assumptions often do not hold and we show, quantitatively and

qualitatively, that they cause errors in the reconstructed 3D shape and pose. To address this, we introduce SPEC, the first in-the-wild 3D HPS method that estimates the perspective camera from a single image and employs this to reconstruct 3D human bodies more accurately. First, we train a neural network to estimate the field of view, camera pitch, and roll given an input image. We employ novel losses that improve the calibration accuracy over previous work. We then train a novel network that concatenates the camera calibration to the image features and uses these together to regress 3D body shape and pose. SPEC is more accurate than the prior art on the standard benchmark (3DPW) as well as two new datasets with more challenging camera views and varying focal lengths. Specifically, we create a new photorealistic synthetic dataset (SPEC-SYN) with ground truth 3D bodies and a novel in-the-wild dataset (SPEC-MTP) with calibration and high-quality reference bodies. Code and datasets are available for research purposes at https://spec.is.tue.mpg.de/.

****

Binocular Mutual Learning for Improving Few-Shot Classification

Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, Chi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8402-8411

Most of the few-shot learning methods learn to transfer knowledge from datasets with abundant labeled data (i.e., the base set). From the perspective of class space on base set, existing methods either focus on utilizing all classes under a global view by normal pretraining, or pay more attention to adopt an episodic manner to train meta-tasks within few classes in a local view. However, the interaction of the two views is rarely explored. As the two views capture complementary information, we naturally think of the compatibility of them for achieving further performance gains. Inspired by the mutual learning paradigm and binocular parallax, we propose a unified framework, namely Binocular Mutual Learning (BML), which achieves the compatibility of the global view and the local view through both intra-view and cross-view modeling. Concretely, the global view learns in the whole class space to capture rich inter-class relationships. Meanwhile, the local view learns in the local class space within each episode, focusing on matching positive pairs correctly. In addition, cross-view mutual interaction further promotes the collaborative learning and the implicit exploration of useful knowledge from each other. During meta-test, binocular embeddings are aggregated together to support decision-making, which greatly improve the accuracy of classification. Extensive experiments conducted on multiple benchmarks including cross-domain validation confirm the effectiveness of our method.

****

Distilling Holistic Knowledge With Graph Neural Networks

Sheng Zhou, Yucheng Wang, Defang Chen, Jiawei Chen, Xin Wang, Can Wang, Jiajun Bu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10387-10396

Knowledge Distillation (KD) aims at transferring knowledge from a larger well-optimized teacher network to a smaller learnable student network. Existing KD methods have mainly considered two types of knowledge, namely the individual knowledge and the relational knowledge. However, these two types of knowledge are usually modeled independently while the inherent correlations between them are largely ignored. It is critical for sufficient student network learning to integrate both individual knowledge and relational knowledge while reserving their inherent correlation. In this paper, we propose to distill the novel holistic knowledge based on an attributed graph constructed among instances. The holistic knowledge is represented as a unified graph-based embedding by aggregating individual knowledge from relational neighborhood samples with graph neural networks, the student network is learned by distilling the holistic knowledge in a contrastive manner. Extensive experiments and ablation studies are conducted on benchmark datasets, the results demonstrate the effectiveness of the proposed method. The code has been published in https://github.com/wyc-ruiker/HKD

****

Towards Robustness of Deep Neural Networks via Regularization

Yao Li, Martin Renqiang Min, Thomas Lee, Wenchao Yu, Erik Kruus, Wei Wang, Cho-J

ui Hsieh; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2021, pp. 7496-7505
Recent studies have demonstrated the vulnerability of deep neural networks again st adversarial examples. Inspired by the observation that adversarial examples o ften lie outside the natural image data manifold and the intrinsic dimension of image data is much smaller than its pixel space dimension, we propose to embed h igh-dimensional input images into a low-dimensional space and apply regularizati on on the embedding space to push the adversarial examples back to the manifold. The proposed framework is called Embedding Regularized Classifier (ER-Classifie r), which improves the adversarial robustness of the classifier through embeddin g regularization. Besides improving classification accuracy against adversarial examples, the framework can be combined with detection methods to detect adversa rial examples. Experimental results on several benchmark datasets show that, our proposed framework achieves good performance against strong adversarial attack methods.
****

STEM: An Approach to Multi-Source Domain Adaptation With Guarantees
Van-Anh Nguyen, Tuan Nguyen, Trung Le, Quan Hung Tran, Dinh Phung; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 935 2-9363
Multi-source Domain Adaptation (MSDA) is more practical but challenging than the conventional unsupervised domain adaptation due to the involvement of diverse m ultiple data sources. Two fundamental challenges of MSDA are: (i) how to deal wi th the diversity in the multiple source domains and (ii) how to cope with the da ta shift between the target domain and the source domains. In this paper, to add ress the first challenge, we propose a theoretical-guaranteed approach to combin e domain experts locally trained on its own source domain to achieve a combined multi-source teacher that globally predicts well on the mixture of source domain s. To address the second challenge, we propose to bridge the gap between the tar get domain and the mixture of source domains in the latent space via a generator or feature extractor. Together with bridging the gap in the latent space, we tr ain a student to mimic the predictions of the teacher expert on both source and target examples. In addition, our approach is guaranteed with rigorous theory of fered insightful justifications of how each component influences the transferrin g performance. Extensive experiments conducted on three benchmark datasets show that our proposed method achieves state-of-the-art performances to the best of o ur knowledge.
****

Divide and Contrast: Self-Supervised Learning From Uncurated Data
Yonglong Tian, Olivier J. Hénaff, Aäron van den Oord; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 10063-10074
Self-supervised learning holds promise in leveraging large amounts of unlabeled data, however much of its progress has thus far been limited to highly curated p re-training data such as ImageNet. We explore the effects of contrastive learnin g from larger, less-curated image datasets such as YFCC, and find there is indee d a large difference in the resulting representation quality. We hypothesize tha t this curation gap is due to a shift in the distribution of image classes---whi ch is more diverse and heavy-tailed---resulting in less relevant negative sample s to learn from. We test this hypothesis with a new approach, Divide and Contras t (DnC), which alternates between contrastive learning and clustering-based hard negative mining. When pretrained on less curated datasets, DnC greatly improves the performance of self-supervised learning on downstream tasks, while remainin g competitive with the current state-of-the-art on curated datasets.
****

Parallel Detection-and-Segmentation Learning for Weakly Supervised Instance Segm entation
Yunhang Shen, Liujuan Cao, Zhiwei Chen, Baochang Zhang, Chi Su, Yongjian Wu, Fei yue Huang, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8198-8208
Weakly supervised instance segmentation (WSIS) with only image-level labels has

recently drawn much attention. To date, bottom-up WSIS methods refine discrimina tive cues from classifiers with sophisticated multi-stage training procedures, w hich also suffer from inconsistent object boundaries. And top-down WSIS methods are formulated as cascade detection-to-segmentation pipeline, in which the quali ty of segmentation learning heavily depends on pseudo masks generated from detec tors. In this paper, we propose a unified parallel detection-and-segmentation le arning (PDSL) framework to learn instance segmentation with only image-level lab els, which draws inspiration from both top-down and bottom-up instance segmentat ion approaches. The detection module is the same as the typical design of any we akly supervised object detection, while the segmentation module leverages self-s upervised learning to model class-agnostic foreground extraction, following by s elf-training to refine class-specific segmentation. We further design instance-a ctivation correlation module to improve the coherence between detection and segm entation branches. Extensive experiments verify that the proposed method outperf orms baselines and achieves the state-of-the-art results on PASCAL VOC and MS CO CO.

****

IntraTomo: Self-Supervised Learning-Based Tomography via Sinogram Synthesis and Prediction

Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, Wolfgang Heidrich; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1960-1970

We propose IntraTomo, a powerful framework that combines the benefits of learnin g-based and model-based approaches for solving highly ill-posed inverse problems in the Computed Tomography (CT) context. IntraTomo is composed of two core modu les: a novel sinogram prediction module, and a geometry refinement module, which are applied iteratively. In the first module, the unknown density field is repr esented as a continuous and differentiable function, parameterized by a deep neu ral network. This network is learned, in a self-supervised fashion, from the inc omplete or/and degraded input sinogram. After getting estimated through the sino gram prediction module, the density field is consistently refined in the second module using local and non-local geometrical priors. With these two core modules , we show that IntraTomo significantly outperforms existing approaches on severa l ill-posed inverse problems, such as limited angle tomography with a range of 4 5 degrees, sparse view tomographic reconstruction with as few as eight views, or super-resolution tomography with eight times increased resolution. The experime nts on simulated and real data show that our approach can achieve results of unp recedented quality.

****

Towards Real-World X-Ray Security Inspection: A High-Quality Benchmark and Later al Inhibition Module for Prohibited Items Detection

Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Y uqing Ma, Libo Zhang, Xianglong Liu; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2021, pp. 10923-10932

Prohibited items detection in X-ray images often plays an important role in prot ecting public safety, which often deals with color-monotonous and luster-insuffi cient objects, resulting in unsatisfactory performance. Till now, there have bee n rare studies touching this topic due to the lack of specialized high-quality d atasets. In this work, we first present a High-quality X-ray (HiXray) security i nspection image dataset, which contains 102,928 common prohibited items of 8 cat egories. It is the largest dataset of high quality for prohibited items detectio n, gathered from the real-world airport security inspection and annotated by pro fessional security inspectors. Besides, for accurate prohibited item detection, we further propose the Lateral Inhibition Module (LIM) inspired by the fact that humans recognize these items by ignoring irrelevant information and focusing on identifiable characteristics, especially when objects are overlapped with each other. Specifically, LIM, the elaborately designed flexible additional module, s uppresses the noisy information flowing maximumly by the Bidirectional Propagati on (BP) module and activates the most identifiable charismatic, boundary, from f our directions by Boundary Activation (BA) module. We evaluate our method extens

ively on HiXray and OPIXray and the results demonstrate that it outperforms SOTA detection methods.
****

Differentiable Surface Rendering via Non-Differentiable Sampling
Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, Zhoutong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6088-6097
We present a method for differentiable rendering of 3D surfaces that supports both explicit and implicit representations, provides derivatives at occlusion boundaries, and is fast and simple to implement. The method first samples the surface using non-differentiable rasterization, then applies differentiable, depth-aware point splatting to produce the final image. Our approach requires no differentiable meshing or rasterization steps, making it efficient for large 3D models and applicable to isosurfaces extracted from implicit surface definitions. We demonstrate the effectiveness of our method for implicit-, mesh-, and parametric-surface-based inverse rendering and neural-network training applications. In particular, we show for the first time efficient, differentiable rendering of an isosurface extracted from a neural radiance field (NeRF), and demonstrate surface-based, rather than volume-based, rendering of a NeRF.
****

Distillation-Guided Image Inpainting
Maitreya Suin, Kuldeep Purohit, A. N. Rajagopalan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2481-2490
Image inpainting methods have shown significant improvements by using deep neural networks recently. However, many of these techniques often create distorted structures or blurry inconsistent textures. The problem is rooted in the encoder layers' ineffectiveness in building a complete and faithful embedding of the missing regions from scratch. Existing solutions like course-to-fine, progressive refinement, structural guidance, etc., suffer from huge computational overheads owing to multiple generator networks, limited ability of handcrafted features, and sub-optimal utilization of the information present in the ground truth. We propose a distillation-based approach for inpainting, where we provide direct feature-level supervision while training. We deploy cross and self-distillation techniques and design a dedicated completion-block in encoder to produce more accurate encoding of the holes. Next, we demonstrate how an inpainting network's attention module can improve by leveraging a distillation-based attention transfer technique and enhancing coherence by using a pixel-adaptive global-local feature fusion. We conduct extensive evaluations on multiple datasets to validate our method. Along with achieving significant improvements over previous SOTA methods, the proposed approach's effectiveness is also demonstrated through its ability to improve existing inpainting works.
****

Real-Time Instance Segmentation With Discriminative Orientation Maps
Wentao Du, Zhiyu Xiang, Shuya Chen, Chengyu Qiao, Yiman Chen, Tingming Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7314-7323
Although instance segmentation has made considerable advancement over recent years, it's still a challenge to design high accuracy algorithms with real-time performance. In this paper, we propose a real-time instance segmentation framework termed OrienMask. Upon the one-stage object detector YOLOv3, a mask head is added to predict some discriminative orientation maps, which are explicitly defined as spatial offset vectors for both foreground and background pixels. Thanks to the discrimination ability of orientation maps, masks can be recovered without the need for extra foreground segmentation. All instances that match with the same anchor size share a common orientation map. This special sharing strategy reduces the amortized memory utilization for mask predictions but without loss of mask granularity. Given the surviving box predictions after NMS, instance masks can be concurrently constructed from the corresponding orientation maps with low complexity. Owing to the concise design for mask representation and its effective integration with the anchor-based object detector, our method is qualified under

real-time conditions while maintaining competitive accuracy. Experiments on COCO benchmark show that OrienMask achieves 34.8 mask AP at the speed of 42.7 fps evaluated with a single RTX 2080 Ti. Code is available at github.com/duwt/OrienMask.

****

Segmenter: Transformer for Semantic Segmentation

Robin Strudel, Ricardo Garcia, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7262-7272

Image segmentation is often ambiguous at the level of individual image patches and requires contextual information to reach label consensus. In this paper we introduce Segmenter, a transformer model for semantic segmentation. In contrast to convolution-based methods, our approach allows to model global context already at the first layer and throughout the network. We build on the recent Vision Transformer (ViT) and extend it to semantic segmentation. To do so, we rely on the output embeddings corresponding to image patches and obtain class labels from these embeddings with a point-wise linear decoder or a mask transformer decoder. We leverage models pre-trained for image classification and show that we can fine-tune them on moderate sized datasets available for semantic segmentation. The linear decoder allows to obtain excellent results already, but the performance can be further improved by a mask transformer generating class masks. We conduct an extensive ablation study to show the impact of the different parameters, in particular the performance is better for large models and small patch sizes. Segmenter attains excellent results for semantic segmentation. It outperforms the state of the art on both ADE20K and Pascal Context datasets and is competitive on Cityscapes.

****

IDARTS: Interactive Differentiable Architecture Search

Song Xue, Runqi Wang, Baochang Zhang, Tian Wang, Guodong Guo, David Doermann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1163-1172

Differentiable Architecture Search (DARTS) improves the efficiency of architecture search by learning the architecture and network parameters end-to-end. However, the intrinsic relationship between the architecture's parameters is neglected, leading to a sub-optimal optimization process. The reason lies in the fact that the gradient descent method used in DARTS ignores the coupling relationship of the parameters and therefore degrades the optimization. In this paper, we address this issue by formulating DARTS as a bilinear optimization problem and introducing an Interactive Differentiable Architecture Search (IDARTS). We first develop a backtracking backpropagation process, which can decouple the relationships of different kinds of parameters and train them in the same framework. The backtracking method coordinates the training of different parameters that fully explore their interaction and optimize training. We present experiments on the CIFAR10 and ImageNet datasets that demonstrate the efficacy of the IDARTS approach by achieving a top-1 accuracy of 76.52% on ImageNet without additional search cost vs. 75.8% with the state-of-the-art PC-DARTS.

****

AutoSpace: Neural Architecture Search With Less Human Interference

Daquan Zhou, Xiaojie Jin, Xiaochen Lian, Linjie Yang, Yujing Xue, Qibin Hou, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 337-346

Current neural architecture search (NAS) algorithms still require expert knowledge and effort to design a search space for network construction. In this paper, we consider automating the search space design to minimize human interference, which however faces two challenges: the explosive complexity of the exploration space and the expensive computation cost to evaluate the quality of different search spaces. To solve them, we propose a novel differentiable evolutionary framework named AutoSpace, which evolves the search space to an optimal one with following novel techniques: a differentiable fitness scoring function to efficiently evaluate the performance of cells and a reference architecture to speedup the evolution procedure and avoid falling into sub-optimal solutions. The framework is

generic and compatible with additional computational constraints, making it fea
sible to learn specialized search spaces that fit different computational budget
s. With the learned search space, the performance of recent NAS algorithms can b
e improved significantly compared with using manually de-signed spaces. Remarkab
ly, the models generated from the new search space achieve 77.8% top-1 accuracy
on ImageNet under the mobile setting (MAdds<=500M), outperforming previous SOTA
EfficientNet-B0 by 0.7%. https://github.com/zhoudaquan/AutoSpace.git
****

Evolving Search Space for Neural Architecture Search
Yuanzheng Ci, Chen Lin, Ming Sun, Boyu Chen, Hongwen Zhang, Wanli Ouyang; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021,
pp. 6659-6669
Automation of neural architecture design has been a coveted alternative to human
 experts. Various search methods have been proposed aiming to find the optimal a
rchitecture in the search space. One would expect the search results to improve
when the search space grows larger since it would potentially contain more perfo
rmant candidates. Surprisingly, we observe that enlarging search space is unbene
ficial or even detrimental to existing NAS methods such as DARTS, ProxylessNAS,
and SPOS. This counterintuitive phenomenon suggests that enabling existing metho
ds to large search space regimes is non-trivial. However, this problem is less d
iscussed in the literature. We present a Neural Search-space Evolution (NSE) sch
eme, the first neural architecture search scheme designed especially for large s
pace neural architecture search problems. The necessity of a well-designed searc
h space with constrained size is a tacit consent in existing methods, and our NS
E aims at minimizing such necessity. Specifically, the NSE starts with a search
space subset, then evolves the search space by repeating two steps: 1) search an
 optimized space from the search space subset, 2) refill this subset from a larg
e pool of operations that are not traversed. We further extend the flexibility o
f obtainable architectures by introducing a learnable multi-branch setting. With
 the proposed method, we achieve 77.3% top-1 retrain accuracy on ImageNet with 3
33M FLOPs, which yielded a state-of-the-art performance among previous auto-gene
rated architectures that do not involve knowledge distillation or weight pruning
. When the latency constraint is adopted, our result also performs better than t
he previous best-performing mobile models with a 77.9% Top-1 retrain accuracy. C
ode is available at https://github.com/orashi/NSE_NAS.
****

THDA: Treasure Hunt Data Augmentation for Semantic Navigation
Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech
Galuba, Stefan Lee, Dhruv Batra; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2021, pp. 15374-15383
Can general-purpose neural models learn to navigate? For PointGoal navigation ("
"go to x, y""), the answer is a clear `yes' -- mapless neural models composed of
 task-agnostic components (CNNs and RNNs) trained with large-scale model-free re
inforcement learning achieve near-perfect performance. However, for ObjectGoal n
avigation (""find a TV""), this is an open question; one we tackle in this paper
. The current best-known result on ObjectNav with general-purpose models is 6% s
uccess rate. First, we show that the key problem is overfitting. Large-scale tra
ining results in 94% success rate on training environments and only 8% in valida
tion. We observe that this stems from agents memorizing environment layouts duri
ng training -- sidestepping the need for exploration and directly learning short
est paths to nearby goal objects. We show that this is a natural consequence of
optimizing for the task metric (which in fact penalizes exploration), is enabled
 by powerful observation encoders, and is possible due to the finite set of trai
ning environment configurations. Informed by our findings, we introduce Treasure
 Hunt Data Augmentation (THDA) to address overfitting in ObjectNav. THDA inserts
 3D scans of household objects at arbitrary scene locations and uses them as Obj
ectNav goals -- augmenting and greatly expanding the set of training layouts. Ta
ken together with our other proposed changes, we improve the state of art on the
 Habitat ObjectGoal Navigation benchmark by 90% (from 14% success rate to 27%) a
nd path efficiency by 48% (from 7.5 SPL to 11.1 SPL).

****
Tripartite Information Mining and Integration for Image Matting
Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, Xin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7555-7564
With the development of deep convolutional neural networks, image matting has ushered in a new phase. Regarding the nature of image matting, most researches have focused on solutions for transition regions. However, we argue that many existing approaches are excessively focused on transition-dominant local fields and ignored the inherent coordination between global information and transition optimisation. In this paper, we propose the Tripartite Information Mining and Integration Network (TIMI-Net) to harmonize the coordination between global and local attributes formally. Specifically, we resort to a novel 3-branch encoder to accomplish comprehensive mining of the input information, which can supplement the neglected coordination between global and local fields. In order to achieve effective and complete interaction between such multi-branches information, we develop the Tripartite Information Integration (TI^2) Module to transform and integrate the interconnections between the different branches. In addition, we built a large-scale human matting dataset (Human-2K) to advance human image matting, which consists of 2100 high-precision human images (2000 images for training and 100 images for test). Finally, we conduct extensive experiments to prove the performance of our proposed TIMI-Net, which demonstrates that our method performs favourably against the SOTA approaches on the alphamatting.com (Rank First), Composition-1K (MSE-0.006, Grad-11.5), Distinctions-646 and our Human-2K. Also, we have developed an online evaluation website to perform natural image matting. Project page: https://wukaoliu.github.io/TIMI-Net.
****
Stochastic Partial Swap: Enhanced Model Generalization and Interpretability for Fine-Grained Recognition
Shaoli Huang, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 620-629
Learning mid-level representation for fine-grained recognition is easily dominated by a limited number of highly discriminative patterns, degrading its robustness and generalization capability. To this end, we propose a novel Stochastic Partial Swap (SPS) scheme to address this issue. Our method performs element-wise swapping for partial features between samples to inject noise during training. It equips a regularization effect similar to Dropout, which promotes more neurons to represent the concepts. Furthermore, it also exhibits other advantages: 1) suppressing over-activation to some part patterns to improve feature representativeness, and 2) enriching pattern combination and simulating noisy cases to enhance classifier generalization. We verify the effectiveness of our approach through comprehensive experiments across four network backbones and three fine-grained datasets. Moreover, we demonstrate its ability to complement high-level representations, allowing a simple model to achieve performance comparable to the top-performing technologies in fine-grained recognition, indoor scene recognition, and material recognition while improving model interpretability.
****
BEV-Net: Assessing Social Distancing Compliance by Joint People Localization and Geometric Reasoning
Zhirui Dai, Yuepeng Jiang, Yi Li, Bo Liu, Antoni B. Chan, Nuno Vasconcelos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5401-5411
ocial distancing, an essential public health measure to limit the spread of contagious diseases, has gained significant attention since the outbreak of the COVID-19 pandemic. In this work, the problem of visual social distancing compliance assessment in busy public areas, with wide field-of-view cameras, is considered. A dataset of crowd scenes with people annotations under a bird's eye view (BEV) and ground truth for metric distances is introduced, and several measures for the evaluation of social distance detection systems are proposed. A multi-branch network, BEV-Net, is proposed to localize individuals in world coordinates and i

dentify high-risk regions where social distancing is violated. BEV-Net combines detection of head and feet locations, camera pose estimation, a differentiable homography module to map image into BEV coordinates, and geometric reasoning to produce a BEV map of the people locations in the scene. Experiments on complex crowded scenes demonstrate the power of the approach and show superior performance over baselines derived from methods in the literature. Applications of interest for public health decision makers are finally discussed. Datasets, code and pretrained models are publicly available at GitHub.
****

Pyramid Architecture Search for Real-Time Image Deblurring
Xiaobin Hu, Wenqi Ren, Kaicheng Yu, Kaihao Zhang, Xiaochun Cao, Wei Liu, Bjoern Menze; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4298-4307
Multi-scale and multi-patch deep models have been shown effective in removing blurs of dynamic scenes. However, these methods still have one major obstacle: manually designing a lightweight and high-efficiency network is challenging and time-consuming. To tackle this problem, we propose a novel deblurring method, dubbed PyNAS (pyramid neural architecture search network), towards automatically designing hyper-parameters including the scales, patches, and standard cell operators. The proposed PyNAS adopts gradient-based search strategies and innovatively searches the hierarchy patch and scale scheme not limited to the cell searching. Specifically, we introduce a hierarchical search strategy tailored for the multi-scale and multi-patch deblurring task. The strategy follows the principle that the first distinguishes between the top-level (pyramid-scales and pyramid-patches) and bottom-level variables (cell operators) and then searches multi-scale variables using the top-to-bottom principle. During the search stage, PyNAS employs an early stopping strategy to avoid the collapse and computational issue. Furthermore, we use a path-level binarization mechanism for multi-scale cell searching to save memory consumption. Our model is a real-time deblurring algorithm (around 58 fps) for 720p images while achieves state-of-the-art deblurring performance on the GoPro and Video Deblurring dataset.
****

TransForensics: Image Forgery Localization With Dense Self-Attention
Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15055-15064
Nowadays advanced image editing tools and technical skills produce tampered images more realistically, which can easily evade image forensic systems and make authenticity verification of images more difficult. To tackle this challenging problem, we introduce TransForensics, a novel image forgery localization method inspired by Transformers. The two major components in our framework are dense self-attention encoders and dense correction modules. The former is to model global context and all pairwise interactions between local patches at different scales, while the latter is used for improving the transparency of the hidden layers and correcting the outputs from different branches. Compared to previous traditional and deep learning methods, TransForensics not only can capture discriminative representations and obtain high-quality mask predictions but is also not limited by tampering types and patch sequence orders. By conducting experiments on main benchmarks, we show that TransForensics outperforms the state-of-the-art methods by a large margin.
****

Joint Audio-Visual Deepfake Detection
Yipin Zhou, Ser-Nam Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14800-14809
Deepfakes ("deep learning" + "fake") are synthetically-generated videos from AI algorithms. While they could be entertaining, they could also be misused for falsifying speeches and spreading misinformation. The process to create deepfakes involves both visual and auditory manipulations. Exploration on detecting visual deepfakes has produced a number of detection methods as well as datasets, while audio deepfakes (e.g. synthetic speech from text-to-speech or voice conversion systems) and the relationship between the visual and auditory modalities have bee

n relatively neglected. In this work, we propose a novel visual / auditory deepf ake joint detection task and show that exploiting the intrinsic synchronization between the visual and auditory modalities could benefit deepfake detection. Exp eriments demonstrate that the proposed joint detection framework outperforms ind ependently trained models, and at the same time, yields superior generalization capability on unseen types of deepfakes.

****

Objects As Cameras: Estimating High-Frequency Illumination From Shadows

Tristan Swedish, Connor Henley, Ramesh Raskar; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2021, pp. 2593-2602

We recover high-frequency information encoded in the shadows cast by an object t o estimate a hemispherical photograph from the viewpoint of the object, effectiv ely turning objects into cameras. Estimating environment maps is useful for adva nced image editing tasks such as relighting, object insertion or removal, and ma terial parameter estimation. Because the problem is ill-posed, recent works in i llumination recovery have tackled the problem of low-frequency lighting for obje ct insertion, rely upon specular surface materials, or make use of data-driven m ethods that are susceptible to hallucination without physically plausible constr aints. We incorporate an optimization scheme to update scene parameters that cou ld enable practical capture of real-world scenes. Furthermore, we develop a meth odology for evaluating expected recovery performance for different types and sha pes of objects.

****

Time-Equivariant Contrastive Video Representation Learning

Simon Jenni, Hailin Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9970-9980

We introduce a novel self-supervised contrastive learning method to learn repres entations from unlabelled videos. Existing approaches ignore the specifics of in put distortions, e.g., by learning invariance to temporal transformations. Inste ad, we argue that video representation should preserve video dynamics and reflec t temporal manipulations of the input. Therefore, we exploit novel constraints t o build representations that are equivariant to temporal transformations and bet ter capture video dynamics. In our method, relative temporal transformations bet ween augmented clips of a video are encoded in a vector and contrasted with othe r transformation vectors. To support temporal equivariance learning, we addition ally propose the self-supervised classification of two clips of a video into 1. overlapping 2. ordered, or 3. unordered. Our experiments show that time-equivari ant representations achieve state-of-the-art results in video retrieval and acti on recognition benchmarks on UCF101, HMDB51, and Diving48.

****

Dynamical Pose Estimation

Heng Yang, Chris Doran, Jean-Jacques Slotine; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2021, pp. 5926-5935

We study the problem of aligning two sets of 3D geometric primitives given known correspondences. Our first contribution is to show that this primitive alignmen t framework unifies five perception problems including point cloud registration, primitive (mesh) registration, category-level 3D registration, absolution pose estimation (APE), and category-level APE. Our second contribution is to propose DynAMical Pose estimation (DAMP), the first general and practical algorithm to s olve primitive alignment problem by simulating rigid body dynamics arising from virtual springs and damping, where the springs span the shortest distances betwe en corresponding primitives. We evaluate DAMP in simulated and real datasets acr oss all five problems, and demonstrate (i) DAMP always converges to the globally optimal solution in the first three problems with 3D-3D correspondences; (ii) a lthough DAMP sometimes converges to suboptimal solutions in the last two problem s with 2D-3D correspondences, using a scheme for escaping local minima, DAMP alw ays succeeds. Our third contribution is to demystify the surprising empirical pe rformance of DAMP and formally prove a global convergence result in the case of point cloud registration by charactering local stability of the equilibrium poin ts of the underlying dynamical system.

****
Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images
Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11148-11157
This paper studies the task of estimating the 3D human poses of multiple persons from multiple calibrated camera views. Following the top-down paradigm, we decompose the task into two stages, i.e. person localization and pose estimation. Both stages are processed in coarse-to-fine manners. And we propose three task-specific graph neural networks for effective message passing. For 3D person localization, we first use Multi-view Matching Graph Module (MMG) to learn the cross-view association and recover coarse human proposals. The Center Refinement Graph Module (CRG) further refines the results via flexible point-based prediction. For 3D pose estimation, the Pose Regression Graph Module (PRG) learns both the multi-view geometry and structural relations between human joints. Our approach achieves state-of-the-art performance on CMU Panoptic and Shelf datasets with significantly lower computation complexity.
****
Learning Fast Sample Re-Weighting Without Reward Data
Zizhao Zhang, Tomas Pfister; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 725-734
Training sample re-weighting is an effective approach for tackling data biases such as imbalanced and corrupted labels. Recent methods develop learning-based algorithms to learn sample re-weighting strategies jointly with model training based on the frameworks of reinforcement learning and meta learning. However, depending on additional unbiased reward data is limiting their general applicability. Furthermore, existing learning-based sample re-weighting methods require nested optimizations of models and weighting parameters, which requires expensive second-order computation. This paper addresses these two problems and presents a novel learning-based fast sample re-weighting (FSR) method that does not require additional reward data. The method is based on two key ideas: learning from history to build proxy reward data and feature sharing to reduce the optimization cost. Our experiments show the proposed method achieves competitive results compared to state of the arts on label noise robustness and long-tailed recognition, and does so while achieving significantly improved training efficiency. The source code is publicly available at https://github.com/google-research/google-research/tree/master/ieg.
****
Multi-Anchor Active Domain Adaptation for Semantic Segmentation
Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, Yefeng Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9112-9122
Unsupervised domain adaption has proven to be an effective approach for alleviating the intensive workload of manual annotation by aligning the synthetic source-domain data and the real-world target-domain samples. Unfortunately, mapping the target-domain distribution to the source-domain unconditionally may distort the essential structural information of the target-domain data. To this end, we firstly propose to introduce a novel multi-anchor based active learning strategy to assist domain adaptation regarding the semantic segmentation task. By innovatively adopting multiple anchors instead of a single centroid, the source domain can be better characterized as a multimodal distribution, thus more representative and complimentary samples are selected from the target domain. With little workload to manually annotate these active samples, the distortion of the target-domain distribution can be effectively alleviated, resulting in a large performance gain. The multi-anchor strategy is additionally employed to model the target-distribution. By regularizing the latent representation of the unlabeled target samples compact around multiple anchors through a novel soft alignment loss, more precise segmentation can be achieved. Extensive experiments are conducted on public datasets to demonstrate that the proposed approach outperforms state-of-the-art methods significantly, along with thorough ablation study to verify the eff

ectiveness of each component.
****
C3-SemiSeg: Contrastive Semi-Supervised Segmentation via Cross-Set Learning and Dynamic Class-Balancing
Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7036-7045
The semi-supervised semantic segmentation methods utilize the unlabeled data to increase the feature discriminative ability to alleviate the burden of the annotated data. However, the dominant consistency learning diagram is limited by a) the misalignment between features from labeled and unlabeled data; b) treating each image and region separately without considering crucial semantic dependencies among classes. In this work, we introduce a novel C^3-SemiSeg to improve consistency-based semi-supervised learning by exploiting better feature alignment under perturbations and enhancing discriminative of the inter-class features cross images. Specifically, we first introduce a cross-set region-level data augmentation strategy to reduce the feature discrepancy between labeled data and unlabeled data. Cross-set pixel-wise contrastive learning is further integrated into the pipeline to facilitate discriminative and consistent intra-class features in a `compared to learn' way. To stabilize training from the noisy label, we propose a dynamic confidence region selection strategy to focus on the high confidence region for loss calculation. We validate the proposed approach on Cityscapes and BDD100K dataset, which significantly outperforms other state-of-the-art semi-supervised semantic segmentation methods.
****
PyMAF: 3D Human Pose and Shape Regression With Pyramidal Mesh Alignment Feedback Loop
Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, Zhenan Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11446-11456
Regression-based methods have recently shown promising results in reconstructing human meshes from monocular images. By directly mapping raw pixels to model parameters, these methods can produce parametric models in a feed-forward manner via a neural networks. However, minor deviation in parameters may lead to noticeable misalignment between the estimated meshes and image evidences. To address this issue, we propose a Pyramidal Mesh Alignment Feedback (PyMAF) loop to leverage a feature pyramid and rectify the predicted parameters explicitly based on the mesh-image alignment status in our deep regressor. In PyMAF, given the currently predicted parameters, mesh-aligned evidences will be extracted from finer-resolution features accordingly and fed back for parameter rectification. To reduce noise and enhance the reliability of these evidences, an auxiliary pixel-wise supervision is imposed on the feature encoder, which provides mesh-image correspondence guidance for our network to preserve the most related information in spatial features. The efficacy of our approach is validated on several benchmarks, including Human3.6M, 3DPW, LSP, and COCO, where experimental results show that our approach consistently improves the mesh-image alignment of the reconstruction. The project page with code and video results can be found at https://hongwenzhang.github.io/pymaf.
****
COOKIE: Contrastive Cross-Modal Knowledge Sharing Pre-Training for Vision-Language Representation
Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, Jie Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2208-2217
There has been a recent surge of interest in cross-modal pre-training. However, existed approaches pre-train a one-stream model to learn joint vision-language representation, which suffers from calculation explosion when conducting cross-modal retrieval. In this work, we propose the Contrastive Cross-Modal Knowledge Sharing Pre-training (COOKIE) method to learn universal text-image representations. There are two key designs in it, one is the weight-sharing transformer on top of the visual and textual encoders to align text and image semantically, the oth

er is three kinds of contrastive learning designed for sharing knowledge between different modalities. Cross-modal knowledge sharing greatly promotes the learning of unimodal representation. Experiments on multi-modal matching tasks including cross-modal retrieval, text matching, and image retrieval show the effectiveness and efficiency of our pre-training framework. Our COOKIE fine-tuned on cross-modal datasets MSCOCO, Flickr30K, and MSRVTT achieves new state-of-the-art results while using only 3/1000 inference time comparing to one-stream models. There are also 5.7 and 3.9 improvements in the task of image retrieval and text matching. Source code will be made public.

****

KoDF: A Large-Scale Korean DeepFake Detection Dataset

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, Gyeongsu Chae; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10744-10753

A variety of effective face-swap and face-reenactment methods have been publicized in recent years, democratizing the face synthesis technology to a great extent. Videos generated as such have come to be called deepfakes with a negative connotation, for various social problems they have caused. Facing the emerging threat of deepfakes, we have built the Korean DeepFake Detection Dataset (KoDF), a large-scale collection of synthesized and real videos focused on Korean subjects. In this paper, we provide a detailed description of methods used to construct the dataset, experimentally show the discrepancy between the distributions of KoDF and existing deepfake detection datasets, and underline the importance of using multiple datasets for real-world generalization. KoDF is publicly available at https://moneybrain-research.github.io/kodf in its entirety (i.e. real clips, synthesized clips, clips with adversarial attack, and metadata).

****

Radial Distortion Invariant Factorization for Structure From Motion

José Pedro Iglesias, Carl Olsson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5906-5915

Factorization methods are frequently used for structure from motion problems (SfM). In the presence of noise they are able to jointly estimate camera matrices and scene points in overdetermined settings, without the need for accurate initial solutions. While the early formulations were restricted to affine models, recent approaches have been show to work with pinhole cameras by minimizing object space errors. In this paper we propose a factorization approach using the so called radial camera, which is invariant to radial distortion and changes in focal length. Assuming a known principal point our approach can reconstruct the 3D scene in settings with unknown and varying radial distortion and focal length. We show on both real and synthetic data that our approach outperforms state-of-the-art factorization methods under these conditions.

****

LaLaLoc: Latent Layout Localisation in Dynamic, Unvisited Environments

Henry Howard-Jenkins, Jose-Raul Ruiz-Sarmiento, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10107-10116

We present LaLaLoc to localise in environments without the need for prior visitation, and in a manner that is robust to large changes in scene appearance, such as a full rearrangement of furniture. Specifically, LaLaLoc performs localisation through latent representations of room layout. LaLaLoc learns a rich embedding space shared between RGB panoramas and layouts inferred from a known floor plan that encodes the structural similarity between locations. Further, LaLaLoc introduces direct, cross-modal pose optimisation in its latent space. Thus, LaLaLoc enables fine-grained pose estimation in a scene without the need for prior visitation, as well as being robust to dynamics, such as a change in furniture configuration. We show that in a domestic environment LaLaLoc is able to accurately localise a single RGB panorama image to within 8.3cm, given only a floor plan as a prior.

****

Learning Privacy-Preserving Optics for Human Pose Estimation

Carlos Hinojosa, Juan Carlos Niebles, Henry Arguello; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2573-2582
The widespread use of always-connected digital cameras in our everyday life has led to increasing concerns about the users' privacy and security. How to develop privacy-preserving computer vision systems? In particular, we want to prevent the camera from obtaining detailed visual data that may contain private information. However, we also want the camera to capture useful information to perform computer vision tasks. Inspired by the trend of jointly designing optics and algorithms, we tackle the problem of privacy-preserving human pose estimation by optimizing an optical encoder (hardware-level protection) with a software decoder (convolutional neural network) in an end-to-end framework. We introduce a visual privacy protection layer in our optical encoder that, parametrized appropriately, enables the optimization of the camera lens's point spread function (PSF). We validate our approach with extensive simulations and a prototype camera. We show that our privacy-preserving deep optics approach successfully degrades or inhibits private attributes while maintaining important features to perform human pose estimation.
****

EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo
Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, Fan Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5732-5740
In this paper, we proposed EPP-MVSNet, a novel deep learning network for 3D reconstruction from multi-view stereo (MVS). EPP-MVSNet can accurately aggregate features at high resolution to a limited cost volume with an optimal depth range, thus, leads to effective and efficient 3D construction. Distinct from existing works which measure feature cost at discrete positions which affects the 3D reconstruction accuracy, EPP-MVSNet introduces an epipolar assembling-based kernel that operates on adaptive intervals along epipolar lines for making full use of the image resolution. Further, we introduce an entropy-based refining strategy where the cost volume describes the space geometry with the little redundancy. Moreover, we design a light-weighted network with Pseudo-3D convolutions integrated to achieve high accuracy and efficiency. We have conducted extensive experiments on challenging datasets Tanks & Temples(TNT), ETH3D and DTU. As a result, we achieve promising results on all datasets and the highest F-Score on the online TNT intermediate benchmark. Code is available at https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/eppmvsnet.
****

Full-Velocity Radar Returns by Radar-Camera Fusion
Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, Praveen Narayanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16198-16207
A distinctive feature of Doppler radar is the measurement of velocity in the radial direction for radar points. However, the missing tangential velocity component hampers object velocity estimation as well as temporal integration of radar sweeps in dynamic scenes. Recognizing that fusing camera with radar provides complementary information to radar, in this paper we present a closed-form solution for the point-wise, full-velocity estimate of Doppler returns using the corresponding optical flow from camera images. Additionally, we address the association problem between radar returns and camera images with a neural network that is trained to estimate radar-camera correspondences. Experimental results on the nuScenes dataset verify the validity of the method and show significant improvements over the state-of-the-art in velocity estimation and accumulation of radar points.
****

Toward Realistic Single-View 3D Object Reconstruction With Unsupervised Learning From Multiple Images
Long-Nhat Ho, Anh Tuan Tran, Quynh Phung, Minh Hoai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12600-12610
Recovering the 3D structure of an object from a single image is a challenging ta

sk due to its ill-posed nature. One approach is to utilize the plentiful photos of the same object category to learn a strong 3D shape prior for the object. Thi s approach has successfully been demonstrated by a recent work of Wu et al. (202 0), which obtained impressive 3D reconstruction networks with unsupervised learn ing. However, their algorithm is only applicable to symmetric objects. In this p aper, we eliminate the symmetry requirement with a novel unsupervised algorithm that can learn a 3D reconstruction network from a multi-image dataset. Our algor ithm is more general and covers the symmetry-required scenario as a special case . Besides, we employ a novel albedo loss that improves the reconstructed details and realisticity. Our method surpasses the previous work in both quality and ro bustness, as shown in experiments on datasets of various structures, including s ingle-view, multi-view, image-collection, and video sets.

****

MVSNeRF: Fast Generalizable Radiance Field Reconstruction From Multi-View Stereo

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, H ao Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14124-14133

We present MVSNeRF, a novel neural rendering approach that can efficiently recon struct neural radiance fields for view synthesis. Unlike prior works on neural r adiance fields that consider per-scene optimization on densely captured images, we propose a generic deep neural network that can reconstruct radiance fields fr om only three nearby input views via fast network inference. Our approach levera ges plane-swept cost volumes (widely used in multi-view stereo) for geometry-awa re scene reasoning, and combines this with physically based volume rendering for neural radiance field reconstruction. We train our network on real objects in t he DTU dataset, and test it on three different datasets to evaluate its effectiv eness and generalizability. Our approach can generalize across scenes (even indo or scenes, completely different from our training scenes of objects) and generat e realistic view synthesis results using only three input images, significantly outperforming concurrent works on generalizable radiance field reconstruction. M oreover, if dense images are captured, our estimated radiance field representati on can be easily fine-tuned; this leads to fast per-scene reconstruction with hi gher rendering quality and substantially less optimization time than NeRF.

****

Transforms Based Tensor Robust PCA: Corrupted Low-Rank Tensors Recovery via Conv ex Optimization

Canyi Lu; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2021, pp. 1145-1152

This work studies the Tensor Robust Principal Component Analysis (TRPCA) problem , which aims to exactly recover the low-rank and sparse components from their su m. Our model is motivated by the recently proposed linear transforms based tenso r-tensor product and tensor SVD. We define a new transforms depended tensor rank and the corresponding tensor nuclear norm. Then we solve the TRPCA problem by c onvex optimization whose objective is a weighted combination of the new tensor n uclear norm and l_1-norm. In theory, we prove that under some incoherence condit ions, the convex program exactly recovers the underlying low-rank and sparse com ponents with high probability. Our new TRPCA is much more general since it allow s to use any invertible linear transforms. Thus, we have more choices in practic e for different tasks and different type of data. Numerical experiments verify o ur results and the application on image recovery demonstrates the superiority of our method.

****

Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin -Brualla, Pratul P. Srinivasan; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 5855-5864

The rendering procedure used by neural radiance fields (NeRF) samples a scene wi th a single ray per pixel and may therefore produce renderings that are excessiv ely blurred or aliased when training or testing images observe scene content at different resolutions. The straightforward solution of supersampling by renderin

g with multiple rays per pixel is impractical for NeRF, because rendering each r ay requires querying a multilayer perceptron hundreds of times. Our solution, wh ich we call "mip-NeRF" (a la "mipmap"), extends NeRF to represent the scene at a continuously-valued scale. By efficiently rendering anti-aliased conical frustu ms instead of rays, mip-NeRF reduces objectionable aliasing artifacts and signif icantly improves NeRF's ability to represent fine details, while also being 7% f aster than NeRF and half the size. Compared to NeRF, mip-NeRF reduces average er ror rates by 17% on the dataset presented with NeRF and by 60% on a challenging multiscale variant of that dataset that we present. Mip-NeRF is also able to mat ch the accuracy of a brute-force supersampled NeRF on our multiscale dataset whi le being 22x faster.

****

Uniformity in Heterogeneity: Diving Deep Into Count Interval Partition for Crowd Counting

Changan Wang, Qingyu Song, Boshen Zhang, Yabiao Wang, Ying Tai, Xuyi Hu, Chengji e Wang, Jilin Li, Jiayi Ma, Yang Wu; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2021, pp. 3234-3242

Recently, the problem of inaccurate learning targets in crowd counting draws inc reasing attention. Inspired by a few pioneering work, we solve this problem by t rying to predict the indices of pre-defined interval bins of counts instead of t he count values themselves. However, an inappropriate interval setting might mak e the count error contributions from different intervals extremely imbalanced, l eading to inferior counting performance. Therefore, we propose a novel count int erval partition criterion called Uniform Error Partition (UEP), which always kee ps the expected counting error contributions equal for all intervals to minimize the prediction risk. Then to mitigate the inevitably introduced discretization errors in the count quantization process, we propose another criterion called Me an Count Proxies (MCP). The MCP criterion selects the best count proxy for each interval to represent its count value during inference, making the overall expec ted discretization error of an image nearly negligible. As far as we are aware, this work is the first to delve into such a classification task and ends up with a promising solution for count interval partition. Following the above two theo retically demonstrated criterions, we propose a simple yet effective model terme d Uniform Error Partition Network (UEPNet), which achieves state-of-the-art perf ormance on several challenging datasets. The codes will be available at: https:/ /github.com/TencentYoutuResearch/CrowdCounting-UEPNet.

****

HDR Video Reconstruction: A Coarse-To-Fine Network and a Real-World Benchmark Da taset

Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K. Wong, Lei Zhan g; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2502-2511

High dynamic range (HDR) video reconstruction from sequences captured with alter nating exposures is a very challenging problem. Existing methods often align low dynamic range (LDR) input sequence in the image space using optical flow, and t hen merge the aligned images to produce HDR output. However, accurate alignment and fusion in the image space are difficult due to the missing details in the ov er-exposed regions and noise in the under-exposed regions, resulting in unpleasi ng ghosting artifacts. To enable more accurate alignment and HDR fusion, we intr oduce a coarse-to-fine deep learning framework for HDR video reconstruction. Fir stly, we perform coarse alignment and pixel blending in the image space to estim ate the coarse HDR video. Secondly, we conduct more sophisticated alignment and temporal fusion in the feature space of the coarse HDR video to produce better r econstruction. Considering the fact that there is no publicly available dataset for quantitative and comprehensive evaluation of HDR video reconstruction method s, we collect such a benchmark dataset, which contains 97 sequences of static sc enes and 184 testing pairs of dynamic scenes. Extensive experiments show that ou r method outperforms previous state-of-the-art methods. Our dataset, code and mo del will be made publicly available.

****

Self Supervision to Distillation for Long-Tailed Visual Recognition
Tianhao Li, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 630-639
Deep learning has achieved remarkable progress for visual recognition on large-scale balanced datasets but still performs poorly on real-world long-tailed data. Previous methods often adopt class re-balanced training strategies to effectively alleviate the imbalance issue, but might be a risk of over-fitting tail classes. The recent decoupling method overcomes over-fitting issues by using a multi-stage training scheme, yet, it is still incapable of capturing tail class information in the feature learning stage. In this paper, we show that soft label can serve as a powerful solution to incorporate label correlation into a multi-stage training scheme for long-tailed recognition. The intrinsic relation between classes embodied by soft labels turns out to be helpful for long-tailed recognition by transferring knowledge from head to tail classes. Specifically, we propose a conceptually simple yet particularly effective multi-stage training scheme, termed as Self Supervised to Distillation (SSD). This scheme is composed of two parts. First, we introduce a self-distillation framework for long-tailed recognition, which can mine the label relation automatically. Second, we present a new distillation label generation module guided by self-supervision. The distilled labels integrate information from both label and data domains that can model long-tailed distribution effectively. We conduct extensive experiments and our method achieves the state-of-the-art results on three long-tailed recognition benchmarks: ImageNet-LT, CIFAR100-LT and iNaturalist 2018. Our SSD outperforms the strong LWS baseline by from 2.7% to 4.5% on various datasets.
****

Learning To Track With Object Permanence
Pavel Tokmakov, Jie Li, Wolfram Burgard, Adrien Gaidon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10860-10869
Tracking by detection, the dominant approach for online multi-object tracking, alternates between localization and association steps. As a result, it strongly depends on the quality of instantaneous observations, often failing when objects are not fully visible. In contrast, tracking in humans is underlined by the notion of object permanence: once an object is recognized, we are aware of its physical existence and can approximately localize it even under full occlusions. In this work, we introduce an end-to-end trainable approach for joint object detection and tracking that is capable of such reasoning. We build on top of the recent CenterTrack architecture, which takes pairs of frames as input, and extend it to videos of arbitrary length. To this end, we augment the model with a spatio-temporal, recurrent memory module, allowing it to reason about object locations and identities in the current frame using all the previous history. It is, however, not obvious how to train such an approach. We study this question on a new, large-scale, synthetic dataset for multi-object tracking, which provides ground truth annotations for invisible objects, and propose several approaches for supervising tracking behind occlusions. Our model, trained jointly on synthetic and real data, outperforms the state of the art on KITTI and MOT17 datasets thanks to its robustness to occlusions.
****

Uncertainty-Guided Transformer Reasoning for Camouflaged Object Detection
Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, Deng-Ping Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4146-4155
Spotting objects that are visually adapted to their surroundings is challenging for both humans and AI. Conventional generic / salient object detection techniques are suboptimal for this task because they tend to only discover easy and clear objects, while overlooking the difficult-to-detect ones with inherent uncertainties derived from indistinguishable textures. In this work, we contribute a novel approach using a probabilistic representational model in combination with transformers to explicitly reason under uncertainties, namely uncertainty-guided transformer reasoning (UGTR), for camouflaged object detection. The core idea is to first learn a conditional distribution over the backbone's output to obtain in

itial estimates and associated uncertainties, and then reason over these uncerta in regions with attention mechanism to produce final predictions. Our approach c ombines the benefits of both Bayesian learning and Transformer-based reasoning, allowing the model to handle camouflaged object detection by leveraging both det erministic and probabilistic information. We empirically demonstrate that our pr oposed approach can achieve higher accuracy than existing state-of-the-art model s on CHAMELEON, CAMO and COD10K datasets. Code is available at https://github.co m/fanyang587/UGTR.
****

Deep Co-Training With Task Decomposition for Semi-Supervised Domain Adaptation
Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q. Weinberger, Wei -Lun Chao, Ser-Nam Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8906-8916
Semi-supervised domain adaptation (SSDA) aims to adapt models trained from a lab eled source domain to a different but related target domain, from which unlabele d data and a small set of labeled data are provided. Current methods that treat source and target supervision without distinction overlook their inherent discre pancy, resulting in a source-dominated model that has not effectively use the ta rget supervision. In this paper, we argue that the labeled target data needs to be distinguished for effective SSDA, and propose to explicitly decompose the SSD A task into two sub-tasks: a semi-supervised learning (SSL) task in the target d omain and an unsupervised domain adaptation (UDA) task across domains. By doing so, the two sub-tasks can better leverage the corresponding supervision and thus yield very different classifiers. To integrate the strengths of the two classif iers, we apply the well established co-training framework, in which the two clas sifiers exchange their high confident predictions to iteratively "teach each oth er" so that both classifiers can excel in the target domain. We call our approac h Deep Co-training with Task decomposition (DeCoTa). DeCoTa requires no adversar ial training and is easy to implement. Moreover, DeCoTa is well founded on the t heoretical condition of when co-training would succeed. As a result, DeCoTa achi eves state-of-the-art results on several SSDA datasets, outperforming the prior art by a notable 4% margin on DomainNet. Code is available at https://github.com /LoyoYang/DeCoTa.
****

Dual Projection Generative Adversarial Networks for Conditional Image Generation
Ligong Han, Martin Renqiang Min, Anastasis Stathopoulos, Yu Tian, Ruijiang Gao, Asim Kadav, Dimitris N. Metaxas; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2021, pp. 14438-14447
Conditional Generative Adversarial Networks (cGANs) extend the standard uncondit ional GAN framework to learning joint data-label distributions from samples, and have been established as powerful generative models capable of generating high- fidelity imagery. A challenge of training such a model lies in properly infusing class information into its generator and discriminator. For the discriminator, class conditioning can be achieved by either (1) directly incorporating labels a s input or (2) involving labels in an auxiliary classification loss. In this pap er, we show that the former directly aligns the class-conditioned fake-and-real data distributions P(\text image |\text class) (data matching), while the la tter aligns data-conditioned class distributions P(\text class |\text image) (label matching). Although class separability does not directly translate to sa mple quality, the discriminator cannot provide useful guidance for the generator if features of distinct classes are mapped to the same point and thus become in separable. Motivated by this intuition, we propose a Dual Projection GAN (P2GAN) model that learns to balance between data matching and label matching . We then propose an improved cGAN model with Auxiliary Classification that directly aligns the fake and real conditionals P(\text class |\text image) by minimizin g their f\mhyphen\text divergence . Experiments on a synthetic Mixture of Gaussi an (MoG) dataset and a variety of real-world datasets including CIFAR100, ImageN et, and VGGFace2 demonstrate the efficacy of our proposed models.
****

EventHPE: Event-Based 3D Human Pose and Shape Estimation

Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun C
hen, Minglun Gong, Li Cheng; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2021, pp. 10996-11005
Event camera is an emerging imaging sensor for capturing dynamics of moving obje
cts as events, which motivates our work in estimating 3D human pose and shape fr
om the event signals. Events, on the other hand, have their unique challenges: r
ather than capturing static body postures, the event signals are best at capturi
ng local motions. This leads us to propose a two-stage deep learning approach, c
alled EventHPE. The first-stage, FlowNet, is trained by unsupervised learning to
 infer optical flow from events. Both events and optical flow are closely relate
d to human body dynamics, which are fed as input to the ShapeNet in the second s
tage, to estimate 3D human shapes. To mitigate the discrepancy between image-bas
ed flow (optical flow) and shape-based flow (vertices movement of human body sha
pe), a novel flow coherence loss is introduced by exploiting the fact that both
flows are originated from the identical human motion. An in-house event-based 3D
 human dataset is curated that comes with 3D pose and shape annotations, which i
s by far the largest one to our knowledge. Empirical evaluations on DHP19 datase
t and our in-house dataset demonstrate the effectiveness of our approach.
****

Synchronization of Group-Labelled Multi-Graphs
Andrea Porfiri Dal Cin, Luca Magri, Federica Arrigoni, Andrea Fusiello, Giacomo
Boracchi; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2021, pp. 6453-6463
Synchronization refers to the problem of inferring the unknown values attached t
o vertices of a graph where edges are labelled with the ratio of the incident ve
rtices, and labels belong to a group. This paper addresses the synchronization p
roblem on multi-graphs, that are graphs with more than one edge connecting the s
ame pair of nodes. The problem naturally arises when multiple measures are avail
able to model the relationship between two vertices. This happens when different
 sensors measure the same quantity, or when the original graph is partitioned in
to sub-graphs that are solved independently. In this case, the relationships amo
ng sub-graphs give rise to multi-edges and the problem can be traced back to a m
ulti-graph synchronization. The baseline solution reduces multi-graphs to simple
 ones by averaging their multi-edges, however this approach falls short because:
 i) averaging is well defined only for some groups and ii) the resulting estimat
or is less precise and accurate, as we prove empirically. Specifically, we prese
nt MultiSynch, a synchronization algorithm for multi-graphs that is based on a p
rincipled constrained eigenvalue optimization. MultiSynch is a general solution
that can cope with any linear group and we show to be profitably usable both on
synthetic and real problems.
****

UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Re
construction
Michael Oechsle, Songyou Peng, Andreas Geiger; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 5589-5599
Neural implicit 3D representations have emerged as a powerful paradigm for recon
structing surfaces from multi-view images and synthesizing novel views. Unfortun
ately, existing methods such as DVR or IDR require accurate per-pixel object mas
ks as supervision. At the same time, neural radiance fields have revolutionized
novel view synthesis. However, NeRF's estimated volume density does not admit ac
curate surface reconstruction. Our key insight is that implicit surface models a
nd radiance fields can be formulated in a unified way, enabling both surface and
 volume rendering using the same model. This unified perspective enables novel,
more efficient sampling procedures and the ability to reconstruct accurate surfa
ces without input masks. We compare our method on the DTU, BlendedMVS, and a syn
thetic indoor dataset. Our experiments demonstrate that we outperform NeRF in te
rms of reconstruction quality while performing on par with IDR without requiring
 masks.
****

CTRL-C: Camera Calibration TRansformer With Line-Classification

Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, Junho Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16228-16237

Single image camera calibration is the task of estimating the camera parameters from a single input image, such as the vanishing points, focal length, and horizon line. In this work, we propose Camera calibration TRansformer with Line-Classification (CTRL-C), an end-to-end neural network-based approach to single image camera calibration, which directly estimates the camera parameters from an image and a set of line segments. Our network adopts the transformer architecture to capture the global structure of an image with multi-modal inputs in an end-to-end manner. We also propose an auxiliary task of line classification to train the network to extract the global geometric information from lines effectively. Our experiments demonstrate that CTRL-C outperforms the previous state-of-the-art methods on the Google Street View and SUN360 benchmark datasets. Code is available at https://github.com/jwlee-vcl/CTRL-C.

****

Parsing Table Structures in the Wild

Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, Gui-Song Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 944-952

This paper tackles the problem of table structure pars-ing (TSP) from images in the wild. In contrast to existingstudies that mainly focus on parsing well-aligned tabularimages with simple layouts from scanned PDF documents,we aim to establish a practical table structure parsing sys-tem for real-world scenarios where tabular input imagesare taken or scanned with severe deformation, bending orocclusions. For designing such a system, we propose anapproach named Cycle-CenterNet on the top of CenterNetwith a novel cycle-pairing module to simultaneously detectand group tabular cells into structured tables. In the cycle-pairing module, a new pairing loss function is proposed forthe network training. Alongside with our Cycle-CenterNet,we also present a large-scale dataset, named Wired Tablein the Wild (WTW), which includes well-annotated structureparsing of multiple style tables in several scenes like photo,scanning files, web pages,etc.. In experiments, we demon-strate that our Cycle-CenterNet consistently achieves thebest accuracy of table structure parsing on the new WTWdataset by 24.6% absolute improvement evaluated by theTEDS metric. A more comprehensive experimental analysisalso validates the advantages of our proposed methods forthe TSP task.

****

Spatio-Temporal Representation Factorization for Video-Based Person Re-Identification

Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, Ziyan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 152-162

Despite much recent progress in video-based person re-identification (re-ID), the current state-of-the-art still suffers from common real-world challenges such as appearance similarity among various people, occlusions, and frame misalignment. To alleviate these problems, we propose Spatio-Temporal Representation Factorization (STRF), a flexible new computational unit that can be used in conjunction with most existing 3D convolutional neural network architectures for re-ID. The key innovations of STRF over prior work include explicit pathways for learning discriminative temporal and spatial features, with each component further factorized to capture complementary person-specific appearance and motion information. Specifically, temporal factorization comprises two branches, one each for static features (e.g., the color of clothes) that do not change much over time, and dynamic features (e.g., walking patterns) that change over time. Further, spatial factorization also comprises two branches to learn both global (coarse segments) as well as local (finer segments) appearance features, with the local features particularly useful in cases of occlusion or spatial misalignment. These two factorization operations taken together result in a modular architecture for our parameter-wise light STRF unit that can be plugged in between any two 3D convolutional layers, resulting in an end-to-end learning framework. We empirically sho

w that STRF improves performance of various existing baseline architectures whil e demonstrating new state-of-the-art results using standard person re-ID evaluat ion protocols on three benchmarks.
****

CondLaneNet: A Top-To-Down Lane Detection Framework Based on Conditional Convolu tion
Lizhe Liu, Xiaohao Chen, Siyu Zhu, Ping Tan; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2021, pp. 3773-3782
Modern deep-learning-based lane detection methods are successful in most scenari os but struggling for lane lines with complex topologies. In this work, we propo se CondLaneNet, a novel top-to-down lane detection framework that detects the la ne instances first and then dynamically predicts the line shape for each instanc e. Aiming to resolve lane instance-level discrimination problem, we introduce a conditional lane detection strategy based on conditional convolution and row-wis e formulation. Further, we design the Recurrent Instance Module(RIM) to overcome the problem of detecting lane lines with complex topologies such as dense lines and fork lines. Benefit from the end-to-end pipeline which requires little post -process, our method has real-time efficiency. We extensively evaluate our metho d on three benchmarks of lane detection. Results show that our method achieves s tate-of-the-art performance on all three benchmark datasets. Moreover, our metho d has the coexistence of accuracy and efficiency, e.g. a 78.14 F1 score and 220 FPS on CULane. Our code is available at https://github.com/aliyun/ conditional-l ane-detection.
****

Adversarial Attacks on Multi-Agent Communication
James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, Raqu el Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 7768-7777
Growing at a fast pace, modern autonomous systems will soon be deployed at scale , opening up the possibility for cooperative multi-agent systems. Sharing inform ation and distributing workloads allow autonomous agents to better perform tasks and increase computation efficiency. However, shared information can be modifie d to execute adversarial attacks on deep learning models that are widely employe d in modern systems. Thus, we aim to study the robustness of such systems and fo cus on exploring adversarial attacks in a novel multi-agent setting where commun ication is done through sharing learned intermediate representations of neural n etworks. We observe that an indistinguishable adversarial message can severely d egrade performance, but becomes weaker as the number of benign agents increases. Furthermore, we show that black-box transfer attacks are more difficult in this setting when compared to directly perturbing the inputs, as it is necessary to align the distribution of learned representations with domain adaptation. Our wo rk studies robustness at the neural network level to contribute an additional la yer of fault tolerance to modern security protocols for more secure multi-agent systems.
****

TransPose: Keypoint Localization via Transformer
Sen Yang, Zhibin Quan, Mu Nie, Wankou Yang; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2021, pp. 11802-11812
While CNN-based models have made remarkable progress on human pose estimation, w hat spatial dependencies they capture to localize keypoints remains unclear. In this work, we propose a model called TransPose, which introduces Transformer for human pose estimation. The attention layers built in Transformer enable our mod el to capture long-range relationships efficiently and also can reveal what depe ndencies the predicted keypoints rely on. To predict keypoint heatmaps, the last attention layer acts as an aggregator, which collects contributions from image clues and forms maximum positions of keypoints. Such a heatmap-based localizatio n approach via Transformer conforms to the principle of Activation Maximization. And the revealed dependencies are image-specific and fine-grained, which also c an provide evidence of how the model handles special cases, e.g., occlusion. The experiments show that TransPose achieves 75.8 AP and 75.0 AP on COCO validation

and test-dev sets, while being more lightweight and faster than mainstream CNN architectures. The TransPose model also transfers very well on MPII benchmark, achieving superior performance on the test set when fine-tuned with small training costs. Code and pre-trained models are publicly available.

****

Vector-Decomposed Disentanglement for Domain-Invariant Object Detection
Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9342-9351

To improve the generalization of detectors, for domain adaptive object detection (DAOD), recent advances mainly explore aligning feature-level distributions between the source and single-target domain, which may neglect the impact of domain-specific information existing in the aligned features. Towards DAOD, it is important to extract domain-invariant object representations. To this end, in this paper, we try to disentangle domain-invariant representations from domain-specific representations. And we propose a novel disentangled method based on vector decomposition. Firstly, an extractor is devised to separate domain-invariant representations from the input, which are used for extracting object proposals. Secondly, domain-specific representations are introduced as the differences between the input and domain-invariant representations. Through the difference operation, the gap between the domain-specific and domain-invariant representations is enlarged, which promotes domain-invariant representations to contain more domain-irrelevant information. In the experiment, we separately evaluate our method on the single- and compound-target case. For the single-target case, experimental results of four domain-shift scenes show our method obtains a significant performance gain over baseline methods. Moreover, for the compound-target case (i.e., the target is a compound of two different domains without domain labels), our method outperforms baseline methods by around 4%, which demonstrates the effectiveness of our method.

****

Topologically Consistent Multi-View Face Inference Using Volumetric Sampling
Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, Yajie Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3824-3834

High-fidelity face digitization solutions often combine multi-view stereo (MVS) techniques for 3D reconstruction and a non-rigid registration step to establish dense correspondence across identities and expressions. A common problem is the need for manual clean-up after the MVS step, as 3D scans are typically affected by noise and outliers and contain hairy surface regions that need to be cleaned up by artists. Furthermore, mesh registration tends to fail for extreme facial expressions. Most learning-based methods use an underlying 3D morphable model (3DMM) to ensure robustness, but this limits the output accuracy for extreme facial expressions. In addition, the global bottleneck of regression architectures cannot produce meshes that tightly fit the ground truth surfaces. We propose ToFu, Topological consistent Face from multi-view, a geometry inference framework that can produce topologically consistent meshes across facial identities and expressions using a volumetric representation instead of an explicit underlying 3DMM. Our novel progressive mesh generation network embeds the topological structure of the face in a feature volume, sampled from geometry-aware local features. A coarse-to-fine architecture facilitates dense and accurate facial mesh predictions in a consistent mesh topology. ToFu further captures displacement maps for pore-level geometric details and facilitates high-quality rendering in the form of albedo and specular reflectance maps. These high-quality assets are readily usable by production studios for avatar creation, animation and physically-based skin rendering. We demonstrate state-of-the-art geometric and correspondence accuracy, while only taking 0.385 seconds to compute a mesh with 10K vertices, which is three orders of magnitude faster than traditional techniques. The code and the model are available for research purposes at https://tianyeli.github.io/tofu.

****

IDM: An Intermediate Domain Module for Domain Adaptive Person Re-ID
Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, Ling-Yu Duan; Proceedin

gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11864-11874
Unsupervised domain adaptive person re-identification (UDA re-ID) aims at transferring the labeled source domain's knowledge to improve the model's discriminability on the unlabeled target domain. From a novel perspective, we argue that the bridging between the source and target domains can be utilized to tackle the UDA re-ID task, and we focus on explicitly modeling appropriate intermediate domains to characterize this bridging. Specifically, we propose an Intermediate Domain Module (IDM) to generate intermediate domains' representations on-the-fly by mixing the source and target domains' hidden representations using two domain factors. Based on the "shortest geodesic path" definition, i.e., the intermediate domains along the shortest geodesic path between the two extreme domains can play a better bridging role, we propose two properties that these intermediate domains should satisfy. To ensure these two properties to better characterize appropriate intermediate domains, we enforce the bridge losses on intermediate domains' prediction space and feature space, and enforce a diversity loss on the two domain factors. The bridge losses aim at guiding the distribution of appropriate intermediate domains to keep the right distance to the source and target domains. The diversity loss serves as a regularization to prevent the generated intermediate domains from being over-fitting to either of the source and target domains. Our proposed method outperforms the state-of-the-arts by a large margin in all the common UDA re-ID tasks, and the mAP gain is up to 7.7% on the challenging MSMT17 benchmark. Code is available at https://github.com/SikaStar/IDM.
****

Robust Watermarking for Deep Neural Networks via Bi-Level Optimization
Peng Yang, Yingjie Lao, Ping Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14841-14850
Deep neural networks (DNNs) have become state-of-the-art in many application domains. The increasing complexity and cost for building these models demand means for protecting their intellectual property (IP). This paper presents a novel DNN framework that optimizes the robustness of the embedded watermarks. Our method is originated from DNN fault attacks. Different from prior end-to-end DNN watermarking approaches, we only modify a tiny subset of weights to embed the watermark, which also facilities better control of the model behaviors and enables larger rooms for optimizing the robustness of the watermarks. In this paper, built upon the above concept, we propose a bi-level optimization framework where the inner loop phase optimizes the example-level problem to generate robust exemplars, while the outer loop phase proposes a masked adaptive optimization to achieve the robustness of the projected DNN models. Our method alternates the learning of the protected models and watermark exemplars across all phases, where watermark exemplars are not just data samples that could be optimized and adjusted instead. We verify the performance of the proposed methods over a wide range of datasets and DNN architectures. Various transformation attacks including fine-tuning, pruning and overwriting are used to evaluate the robustness.
****

Efficient Video Compression via Content-Adaptive Super-Resolution
Mehrdad Khani, Vibhaalakshmi Sivaraman, Mohammad Alizadeh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4521-4530
Video compression is a critical component of Internet video delivery. Recent work has shown that deep learning techniques can rival or outperform human-designed algorithms, but these methods are significantly less compute and power-efficient than existing codecs. This paper presents a new approach that augments existing codecs with a small, content-adaptive super-resolution model that significantly boosts video quality. Our method, SRVC, encodes video into two bitstreams: (i) a content stream, produced by compressing downsampled low-resolution video with the existing codec, (ii) a model stream, which encodes periodic updates to a lightweight super-resolution neural network customized for short segments of the video. SRVC decodes the video by passing the decompressed low-resolution video frames through the (time-varying) super-resolution model to reconstruct high-resolution video frames. Our results show that to achieve the same PSNR, SRVC require

s 20% of the bits-per-pixel of H.265 in slow mode, and 3% of the bits-per-pixel of DVC, a recent deep learning-based video compression scheme. SRVC runs at 90 frames per second on an NVIDIA V100 GPU.
****

Video Annotation for Visual Tracking via Selection and Refinement
Kenan Dai, Jie Zhao, Lijun Wang, Dong Wang, Jianhua Li, Huchuan Lu, Xuesheng Qian, Xiaoyun Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10296-10305
Deep learning based visual trackers entail offline pre-training on large volumes of video datasets with accurate bounding box annotations that are labor-expensive to achieve. We present a new framework to facilitate bounding box annotations for video sequences, which investigates a selection-and-refinement strategy to automatically improve the preliminary annotations generated by tracking algorithms. A temporal assessment network (T-Assess Net) is proposed which is able to capture the temporal coherence of target locations and select reliable tracking results by measuring their quality. Meanwhile, a visual-geometry refinement network (VG-Refine Net) is also designed to further enhance the selected tracking results by considering both target appearance and temporal geometry constraints, allowing inaccurate tracking results to be corrected. The combination of the above two networks provides a principled approach to ensure the quality of automatic video annotation. Experiments on large scale tracking benchmarks demonstrate that our method can deliver highly accurate bounding box annotations and significantly reduce human labor by 94.0%, yielding an effective means to further boost tracking performance with augmented training data.
****

A Unified 3D Human Motion Synthesis Model via Conditional Variational Auto-Encoder
Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, Xiaohui Shen, Ding Liu, Nadia Magnenat Thalmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11645-11655
We present a unified and flexible framework to address the generalized problem of 3D motion synthesis that covers the tasks of motion prediction, completion, interpolation, and spatial-temporal recovery. Since these tasks have different input constraints and various fidelity and diversity requirements, most existing approaches only cater to a specific task or use different architectures to address various tasks. Here we propose a unified framework based on Conditional Variational Auto-Encoder (CVAE), where we treat any arbitrary input as a masked motion series. Notably, by considering this problem as a conditional generation process, we estimate a parametric distribution of the missing regions based on the input conditions, from which to sample and synthesize the full motion series. To further allow the flexibility of manipulating the motion style of the generated series, we design an Action-Adaptive Modulation (AAM) to propagate the given semantic guidance through the whole sequence. We also introduce a cross-attention mechanism to exploit distant relations among decoder and encoder features for better realism and global consistency. We conducted extensive experiments on Human 3.6M and CMU-Mocap. The results show that our method produces coherent and realistic results for various motion synthesis tasks, with the synthesized motions distinctly adapted by the given action labels.
****

SaccadeCam: Adaptive Visual Attention for Monocular Depth Sensing
Brevin Tilmon, Sanjeev J. Koppal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6009-6018
Most monocular depth sensing methods use conventionally captured images that are created without considering scene content. In contrast, animal eyes have fast mechanical motions, called saccades, that control how the scene is imaged by the fovea, where resolution is highest. In this paper, we present the SaccadeCam framework for adaptively distributing resolution onto regions of interest in the scene. Our algorithm for adaptive resolution is a self-supervised network and we demonstrate results for end-to-end learning for monocular depth estimation. We al

so show preliminary results with a real SaccadeCam hardware prototype.
****

Safety-Aware Motion Prediction With Unseen Vehicles for Autonomous Driving
Xuanchi Ren, Tao Yang, Li Erran Li, Alexandre Alahi, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15731-15740
Motion prediction of vehicles is critical but challenging due to the uncertainties in complex environments and the limited visibility caused by occlusions and limited sensor ranges. In this paper, we study a new task, safety-aware motion prediction with unseen vehicles for autonomous driving. Unlike the existing trajectory prediction task for seen vehicles, we aim at predicting an occupancy map that indicates the earliest time when each location can be occupied by either seen and unseen vehicles. The ability to predict unseen vehicles is critical for safety in autonomous driving. To tackle this challenging task, we propose a safety-aware deep learning model with three new loss functions to predict the earliest occupancy map. Experiments on the large-scale autonomous driving nuScenes dataset show that our proposed model significantly outperforms the state-of-the-art baselines on the safety-aware motion prediction task. To the best of our knowledge, our approach is the first one that can predict the existence of unseen vehicles in most cases.
****

Mesh Graphormer
Kevin Lin, Lijuan Wang, Zicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12939-12948
We present a graph-convolution-reinforced transformer, named Mesh Graphormer, for 3D human pose and mesh reconstruction from a single image. Recently both transformers and graph convolutional neural networks (GCNNs) have shown promising progress in human mesh reconstruction. Transformer-based approaches are effective in modeling non-local interactions among 3D mesh vertices and body joints, whereas GCNNs are good at exploiting neighborhood vertex interactions based on a pre-specified mesh topology. In this paper, we study how to combine graph convolutions and self-attentions in a transformer to model both local and global interactions. Experimental results show that our proposed method, Mesh Graphormer, significantly outperforms the previous state-of-the-art methods on multiple benchmarks, including Human3.6M, 3DPW, and FreiHAND datasets. Code and pre-trained models are available at https://github.com/microsoft/MeshGraphormer
****

CrossNorm and SelfNorm for Generalization Under Distribution Shifts
Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, Dimitris N. Metaxas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 52-61
Traditional normalization techniques (e.g., Batch Normalization and Instance Normalization) generally and simplistically assume that training and test data follow the same distribution. As distribution shifts are inevitable in real-world applications, well-trained models with previous normalization methods can perform badly in new environments. Can we develop new normalization methods to improve generalization robustness under distribution shifts? In this paper, we answer the question by proposing CrossNorm and SelfNorm. CrossNorm exchanges channel-wise mean and variance between feature maps to enlarge training distribution, while SelfNorm uses attention to recalibrate the statistics to bridge gaps between training and test distributions. CrossNorm and SelfNorm can complement each other, though exploring different directions in statistics usage. Extensive experiments on different fields (vision and language), tasks (classification and segmentation), settings (supervised and semi-supervised), and distribution shift types (synthetic and natural) show the effectiveness. Code is available at https://github.com/amazon-research/crossnorm-selfnorm
****

Elaborative Rehearsal for Zero-Shot Action Recognition
Shizhe Chen, Dong Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13638-13647

The growing number of action classes has posed a new challenge for video understanding, making Zero-Shot Action Recognition (ZSAR) a thriving direction. The ZSAR task aims to recognize target (unseen) actions without training examples by leveraging semantic representations to bridge seen and unseen actions. However, due to the complexity and diversity of actions, it remains challenging to semantically represent action classes and transfer knowledge from seen data. In this work, we propose an ER-enhanced ZSAR model inspired by an effective human memory technique Elaborative Rehearsal (ER), which involves elaborating a new concept and relating it to known concepts. Specifically, we expand each action class as an Elaborative Description (ED) sentence, which is more discriminative than a class name and less costly than manual-defined attributes. Besides directly aligning class semantics with videos, we incorporate objects from the video as Elaborative Concepts (EC) to improve video semantics and generalization from seen actions to unseen actions. Our ER-enhanced ZSAR model achieves state-of-the-art results on three existing benchmarks. Moreover, we propose a new ZSAR evaluation protocol on the Kinetics dataset to overcome limitations of current benchmarks and first compare with few-shot learning baselines on this more realistic setting. Our codes and collected EDs are released at https://github.com/DeLightCMU/ElaborativeRehearsal.

****

CAG-QIL: Context-Aware Actionness Grouping via Q Imitation Learning for Online Temporal Action Localization

Hyolim Kang, Kyungmin Kim, Yumin Ko, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13729-13738

Temporal action localization has been one of the most popular tasks in video understanding, due to the importance of detecting action instances in videos. However, not much progress has been made on extending it to work in an online fashion, although many video related tasks can benefit by going online with the growing video streaming services. To this end, we introduce a new task called Online Temporal Action Localization (On-TAL), in which the goal is to immediately detect action instances from an untrimmed streaming video. The online setting makes the new task very challenging as the actionness decision for every frame has to be made without access to future frames and also because post-processing methods cannot be used to modify past action proposals. We propose a novel framework, Context-Aware Actionness Grouping (CAG) as a solution for On-TAL and train it with the imitation learning algorithm, which allows us to avoid sophisticated reward engineering. Evaluation of our work on THUMOS14 and Activitynet1.3 shows significant improvement over non-naive baselines, demonstrating the effectiveness of our approach. As a by-product, our method can also be used for the Online Detection of Action Start (ODAS), in which our method also outperforms previous state-of-the-art models.

****

Joint Visual and Audio Learning for Video Highlight Detection

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, Li Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8127-8137

In video highlight detection, the goal is to identify the interesting moments within an unedited video. Although the audio component of the video provides important cues for highlight detection, the majority of existing efforts focus almost exclusively on the visual component. In this paper, we argue that both audio and visual components of a video should be modeled jointly to retrieve its best moments. To this end, we propose an audio-visual network for video highlight detection. At the core of our approach lies a bimodal attention mechanism, which captures the interaction between the audio and visual components of a video, and produces fused representations to facilitate highlight detection. Furthermore, we introduce a noise sentinel technique to adaptively discount a noisy visual or audio modality. Empirical evaluations on two benchmark datasets demonstrate the superior performance of our approach over the state-of-the-art methods.

****

Shallow Bayesian Meta Learning for Real-World Few-Shot Recognition

Xueting Zhang, Debin Meng, Henry Gouk, Timothy M. Hospedales; Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 651-660
Many state-of-the-art few-shot learners focus on developing effective training procedures for feature representations, before using simple (e.g., nearest centroid) classifiers. We take an approach that is agnostic to the features used, and focus exclusively on meta-learning the final classifier layer. Specifically, we introduce MetaQDA, a Bayesian meta-learning generalisation of the classic quadratic discriminant analysis. This approach has several benefits of interest to practitioners: meta-learning is fast and memory efficient, without the need to fine-tune features. It is agnostic to the off-the-shelf features chosen, and thus will continue to benefit from future advances in feature representations. Empirically, it leads to excellent performance in cross-domain few-shot learning, class-incremental few-shot learning, and crucially for real-world applications, the Bayesian formulation leads to state-of-the-art uncertainty calibration in predictions.
****

Towards Interpretable Deep Metric Learning With Structural Matching
Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9887-9896
How do the neural networks distinguish two images? It is of critical importance to understand the matching mechanism of deep models for developing reliable intelligent systems for many risky visual applications such as surveillance and access control. However, most existing deep metric learning methods match the images by comparing feature vectors, which ignores the spatial structure of images and thus lacks interpretability. In this paper, we present a deep interpretable metric learning (DIML) method for more transparent embedding learning. Unlike conventional metric learning methods based on feature vector comparison, we propose a structural matching strategy that explicitly aligns the spatial embeddings by computing an optimal matching flow between feature maps of the two images. Our method enables deep models to learn metrics in a more human-friendly way, where the similarity of two images can be decomposed to several part-wise similarities and their contributions to the overall similarity. Our method is model-agnostic, which can be applied to off-the-shelf backbone networks and metric learning methods. We evaluate our method on three major benchmarks of deep metric learning including CUB200- 2011, Cars196, and Stanford Online Products, and achieve substantial improvements over popular metric learning methods with better interpretability.
****

Weakly Supervised Text-Based Person Re-Identification
Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, Nong Sang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11395-11404
The conventional text-based person re-identification methods heavily rely on identity annotations. However, this labeling process is costly and time-consuming. In this paper, we consider a more practical setting called weakly supervised text-based person re-identification, where only the text-image pairs are available without the requirement of annotating identities during the training phase. To this end, we propose a Cross-Modal Mutual Training (CMMT) framework. Specifically, to alleviate the intra-class variations, a clustering method is utilized to generate pseudo labels for both visual and textual instances. To further refine the clustering results, CMMT provides a Mutual Pseudo Label Refinement module, which leverages the clustering results in one modality to refine that in the other modality constrained by the text-image pairwise relationship. Meanwhile, CMMT introduces a Text-IoU Guided Cross-Modal Projection Matching loss to resolve the cross-modal matching ambiguity problem. A Text-IoU Guided Hard Sample Mining method is also proposed for learning discriminative textual-visual joint embeddings. We conduct extensive experiments to demonstrate the effectiveness of the proposed CMMT, and the results show that CMMT performs favorably against existing text-based person re-identification methods.
****

Learning Temporal Dynamics From Cycles in Narrated Video

Dave Epstein, Jiajun Wu, Cordelia Schmid, Chen Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1480-1489

Learning to model how the world changes as time elapses has proven a challenging problem for the computer vision community. We introduce a self-supervised approach to this problem that solves a multi-modal temporal cycle consistency objective, MMCC, jointly in vision and language. This objective requires a model to learn modality-agnostic functions to predict the future and past that undo each other when composed. We hypothesize that a model trained on this objective will discover long-term temporal dynamics in video. We verify this hypothesis by using the resultant visual representations and predictive models as-is to solve a variety of downstream tasks. Our method outperforms state-of-the-art self-supervised video prediction methods on future action anticipation, temporal image ordering, and arrow-of-time classification tasks, without training on target datasets or their labels.

****

von Mises-Fisher Loss: An Exploration of Embedding Geometries for Supervised Learning

Tyler R. Scott, Andrew C. Gallagher, Michael C. Mozer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10612-10622

Recent work has argued that classification losses utilizing softmax cross-entropy are superior not only for fixed-set classification tasks, but also by outperforming losses developed specifically for open-set tasks including few-shot learning and retrieval. Softmax classifiers have been studied using different embedding geometries---Euclidean, hyperbolic, and spherical---and claims have been made about the superiority of one or another, but they have not been systematically compared with careful controls. We conduct an empirical investigation of embedding geometry on softmax losses for a variety of fixed-set classification and image retrieval tasks. An interesting property observed for the spherical losses leads us to propose a probabilistic classifier based on the von Mises-Fisher distribution, and we show that it is competitive with state-of-the-art methods while producing improved out-of-the-box calibration. We provide guidance regarding the trade-offs between losses and how to choose among them.

****

Multiple Heads Are Better Than One: Few-Shot Font Generation With Multiple Localized Experts

Song Park, Sanghyuk Chun, Junbum Cha, Bado Lee, Hyunjung Shim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13900-13909

A few-shot font generation (FFG) method has to satisfy two objectives: the generated images should preserve the underlying global structure of the target character and present the diverse local reference style. Existing FFG methods aim to disentangle content and style either by extracting a universal representation style or extracting multiple component-wise style representations. However, previous methods either fail to capture diverse local styles or cannot be generalized to a character with unseen components, e.g., unseen language systems. To mitigate the issues, we propose a novel FFG method, named Multiple Localized Experts Few-shot Font Generation Network (MX-Font). MX-Font extracts multiple style features not explicitly conditioned on component labels, but automatically by multiple experts to represent different local concepts, e.g., left-side sub-glyph. Owing to the multiple experts, MX-Font can capture diverse local concepts and show the generalizability to unseen languages. During training, we utilize component labels as weak supervision to guide each expert to be specialized for different local concepts. We formulate the component assign problem to each expert as the graph matching problem, and solve it by the Hungarian algorithm. We also employ the independence loss and the content-style adversarial loss to impose the content-style disentanglement. In our experiments, MX-Font outperforms previous state-of-the-art FFG methods in the Chinese generation and cross-lingual, e.g., Chinese to Korean, generation.

****

Me-Momentum: Extracting Hard Confident Examples From Noisily Labeled Data

Yingbin Bai, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9312-9321

Examples that are close to the decision boundary---that we term hard examples, are essential to shape accurate classifiers. Extracting confident examples has been widely studied in the community of learning with noisy labels. However, it remains elusive how to extract hard confident examples from the noisy training data. In this paper, we propose a deep learning paradigm to solve this problem, which is built on the memorization effect of deep neural networks that they would first learn simple patterns, i.e., which are defined by these shared by multiple training examples. To extract hard confident examples that contain non-simple patterns and are entangled with the inaccurately labeled examples, we borrow the idea of momentum from physics. Specifically, we alternately update the confident examples and refine the classifier. Note that the extracted confident examples in the previous round can be exploited to learn a better classifier and that the better classifier will help identify better (and hard) confident examples. We call the approach the "Momentum of Memorization" (Me-Momentum). Empirical results on benchmark-simulated and real-world label-noise data illustrate the effectiveness of Me-Momentum for extracting hard confident examples, leading to better classification performance.

****

mDALU: Multi-Source Domain Adaptation and Label Unification With Partial Datasets

Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8876-8885

One challenge of object recognition is to generalize to new domains, to more classes and/or to new modalities. This necessitates methods to combine and reuse existing datasets that may belong to different domains, have partial annotations, and/or have different data modalities. This paper formulates this as a multi-source domain adaptation and label unification problem, and proposes a novel method for it. Our method consists of a partially-supervised adaptation stage and a fully-supervised adaptation stage. In the former, partial knowledge is transferred from multiple source domains to the target domain and fused therein. Negative transfer between unmatching label spaces is mitigated via three new modules: domain attention, uncertainty maximization and attention-guided adversarial alignment. In the latter, knowledge is transferred in the unified label space after a label completion process with pseudo-labels. Extensive experiments on three different tasks - image classification, 2D semantic image segmentation, and joint 2D-3D semantic segmentation - show that our method outperforms all competing methods significantly.

****

Collaging Class-Specific GANs for Semantic Image Synthesis

Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, Krishna Kumar Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14418-14427

We propose a new approach for high resolution semantic image synthesis. It consists of one base image generator and multiple class-specific generators. The base generator generates high quality images based on a segmentation map. To further improve the quality of different objects, we create a bank of Generative Adversarial Networks (GANs) by separately training class-specific models. This has several benefits including -- dedicated weights for each class; centrally aligned data for each model; additional training data from other sources, potential of higher resolution and quality; and easy manipulation of a specific object in the scene. Experiments show that our approach can generate high quality images in high resolution while having flexibility of object-level control by using class-specific generators.

****

Meta Navigator: Search for a Good Adaptation Policy for Few-Shot Learning

Chi Zhang, Henghui Ding, Guosheng Lin, Ruibo Li, Changhu Wang, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9435-9444

Few-shot learning aims to adapt knowledge learned from previous tasks to novel t
asks with only a limited amount of labeled data. Research literature on few-shot
 learning exhibits great diversity, while different algorithms often excel at di
fferent few-shot learning scenarios. It is therefore tricky to decide which lear
ning strategies to use under different task conditions. Inspired by the recent s
uccess in Automated Machine Learning literature (AutoML), in this paper, we pres
ent Meta Navigator, a framework that attempts to solve the aforementioned limita
tion in few-shot learning by seeking a higher-level strategy and proffer to auto
mate the selection from various few-shot learning designs. The goal of our work
 is to search for good parameter adaptation policies that are applied to differen
t stages in the network for few-shot classification. We present a search space t
hat covers many popular few-shot learning algorithms in the literature and devel
op a differentiable searching and decoding algorithm based on meta-learning that
 supports gradient-based optimization. We demonstrate the effectiveness of our s
earching-based method on multiple benchmark datasets. Extensive experiments show
 that our approach significantly outperforms baselines and demonstrates performa
nce advantages over many state-of-the-art methods. Code and models will be made
publicly available.
****

Occlusion-Aware Video Object Inpainting
Lei Ke, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2021, pp. 14468-14478
Conventional video inpainting is neither object-oriented nor occlusion-aware, ma
king it liable to obvious artifacts when large occluded object regions are inpai
nted. This paper presents occlusion-aware video object inpainting, which recover
s both the complete shape and appearance for occluded objects in videos given th
eir visible mask segmentation. To facilitate this new research, we construct the
 first large-scale video object inpainting benchmark YouTube-VOI to provide r
ealistic occlusion scenarios with both occluded and visible object masks availab
le. Our technical contribution VOIN jointly performs video object shape completi
on and occluded texture generation. In particular, the shape completion module m
odels long-range object coherence while the flow completion module recovers accu
rate flow with sharp motion boundary, for propagating temporally-consistent text
ure to the same moving object across frames. For more realistic results, VOIN is
 optimized using both T-PatchGAN and a new spatio-temporal attention-based multi
-class discriminator. Finally, we compare VOIN and strong baselines on YouTube-V
OI. Experimental results clearly demonstrate the efficacy of our method includin
g inpainting complex and dynamic objects. VOIN degrades gracefully with inaccura
te input visible mask.
****

TransFER: Learning Relation-Aware Facial Expression Representations With Transfo
rmers
Fanglei Xue, Qiangchang Wang, Guodong Guo; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 3601-3610
Facial expression recognition (FER) has received increasing interest in computer
 vision. We propose the TransFER model which can learn rich relation-aware local
 representations. It mainly consists of three components: Multi-Attention Droppi
ng (MAD), ViT-FER, and Multi-head Self-Attention Dropping (MSAD). First, local p
atches play an important role in distinguishing various expressions, however, fe
w existing works can locate discriminative and diverse local patches. This can c
ause serious problems when some patches are invisible due to pose variations or
viewpoint changes. To address this issue, the MAD is proposed to randomly drop a
n attention map. Consequently, models are pushed to explore diverse local patche
s adaptively. Second, to build rich relations between different local patches, t
he Vision Transformers (ViT) are used in FER, called ViT-FER. Since the global s
cope is used to reinforce each local patch, a better representation is obtained
to boost the FER performance. Thirdly, the multi-head self-attention allows ViT
to jointly attend to features from different information subspaces at different
positions. Given no explicit guidance, however, multiple self-attentions may ext
ract similar relations. To address this, the MSAD is proposed to randomly drop o

ne self-attention module. As a result, models are forced to learn rich relations among diverse local patches. Our proposed TransFER model outperforms the state-of-the-art methods on several FER benchmarks, showing its effectiveness and usefulness.

****

Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval
Frederik Warburg, Martin Jørgensen, Javier Civera, Søren Hauberg; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12158-12168

Uncertainty quantification in image retrieval is crucial for downstream decisions, yet it remains a challenging and largely unexplored problem. Current methods for estimating uncertainties are poorly calibrated, computationally expensive, or based on heuristics. We present a new method that views image embeddings as stochastic features rather than deterministic features. Our two main contributions are (1) a likelihood that matches the triplet constraint and that evaluates the probability of an anchor being closer to a positive than a negative; and (2) a prior over the feature space that justifies the conventional l2 normalization. To ensure computational efficiency, we derive a variational approximation of the posterior, called the Bayesian triplet loss, that produces state-of-the-art uncertainty estimates and matches the predictive performance of current state-of-the-art methods.

****

Manifold Matching via Deep Metric Learning for Generative Modeling
Mengyu Dai, Haibin Hang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6587-6597

We propose a manifold matching approach to generative models which includes a distribution generator (or data generator) and a metric generator. In our framework, we view the real data set as some manifold embedded in a high-dimensional Euclidean space. The distribution generator aims at generating samples that follow some distribution condensed around the real data manifold. It is achieved by matching two sets of points using their geometric shape descriptors, such as centroid and p-diameter, with learned distance metric; the metric generator utilizes both real data and generated samples to learn a distance metric which is close to some intrinsic geodesic distance on the real data manifold. The produced distance metric is further used for manifold matching. The two networks learn simultaneously during the training process. We apply the approach on both unsupervised and supervised learning tasks: in unconditional image generation task, the proposed method obtains competitive results compared with existing generative models; in super-resolution task, we incorporate the framework in perception-based models and improve visual qualities by producing samples with more natural textures. Experiments and analysis demonstrate the feasibility and effectiveness of the proposed framework.

****

ProFlip: Targeted Trojan Attack With Progressive Bit Flips
Huili Chen, Cheng Fu, Jishen Zhao, Farinaz Koushanfar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7718-7727

The security of Deep Neural Networks (DNNs) is of great importance due to their employment in various safety-critical applications. DNNs are shown to be vulnerable against the Trojan attack that manipulates the model parameters via poisoned training and gets activated by the pre-defined trigger in inputs during inference. In this work, we present ProFlip, the first targeted Trojan attack framework that can divert the prediction of the DNN to the target class by progressively identifying and flipping a small set of bits in model parameters. At its core, ProFlip consists of three key phases: (i) Determining significant neurons in the last layer; (ii) Generating an effective trigger pattern for the target class; (iii) Identifying a sequence of susceptible bits of DNN parameters stored in the main memory (e.g., DRAM). After model deployment, the adversary can insert the Trojan by flipping the critical bits found by ProFlip using bit flip techniques such as Row Hammer or laser beams. As the result, the altered DNN predicts the target class when the trigger pattern is present in any inputs. We perform extensi

ve evaluations of ProFlip on CIFAR10, SVHN, and ImageNet datasets with ResNet-18 and VGG-16 architectures. Empirical results show that, to reach an attack success rate (ASR) of over 94%, ProFlip requires only 12 bit flips out of 88 million parameter bits for ResNet-18 with CIFAR-10, and 15 bit flips for ResNet-18 with ImageNet. Compared to the SOTA, ProFlip reduces the number of required bits flips by 28x 34x while reaching the same level of ASR.
****

AutoFormer: Searching Transformers for Visual Recognition
Minghao Chen, Houwen Peng, Jianlong Fu, Haibin Ling; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12270-12280
Recently, pure transformer-based models have shown great potentials for vision tasks such as image classification and detection. However, the design of transformer networks is challenging. It has been observed that the depth, embedding dimension, and number of heads can largely affect the performance of vision transformers. Previous models configure these dimensions based upon manual crafting. In this work, we propose a new one-shot architecture search framework, namely AutoFormer, dedicated to vision transformer search. AutoFormer entangles the weights of different blocks in the same layers during supernet training. Benefiting from the strategy, the trained supernet allows thousands of subnets to be very well-trained. Specifically, the performance of these subnets with weights inherited from the supernet is comparable to those retrained from scratch. Besides, the searched models, which we refer to AutoFormers, surpass the recent state-of-the-arts such as ViT and DeiT. In particular, AutoFormer-tiny/small/base achieve 74.7%/81.7%/82.4% top-1 accuracy on ImageNet with 5.7M/22.9M/53.7M parameters, respectively. Lastly, we verify the transferability of AutoFormer by providing the performance on downstream benchmarks and distillation experiments. Code and models are available at https://github.com/microsoft/Cream.
****

Mining Latent Classes for Few-Shot Segmentation
Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8721-8730
Few-shot segmentation (FSS) aims to segment unseen classes given only a few annotated samples. Existing methods suffer the problem of feature undermining, i.e. potential novel classes are treated as background during training phase. Our method aims to alleviate this problem and enhance the feature embedding on latent novel classes. In our work, we propose a novel joint-training framework. Based on conventional episodic training on support-query pairs, we add an additional mining branch that exploits latent novel classes via transferable sub-clusters, and a new rectification technique on both background and foreground categories to enforce more stable prototypes. Over and above that, our transferable sub-cluster has the ability to leverage extra unlabeled data for further feature enhancement. Extensive experiments on two FSS benchmarks demonstrate that our method outperforms previous state-of-the-art by a large margin of 3.7% mIOU on PASCAL-5i and 7.0% mIOU on COCO-20i at the cost of 74% fewer parameters and 2.5x faster inference speed. The source code is available at https://github.com/LiheYoung/MiningFSS.
****

Active Learning for Deep Object Detection via Probabilistic Modeling
Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, Jose M. Alvarez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10264-10273
Active learning aims to reduce labeling costs by selecting only the most informative samples on a dataset. Few existing works have addressed active learning for object detection. Most of these methods are based on multiple models or are straightforward extensions of classification methods, hence estimate an image's informativeness using only the classification head. In this paper, we propose a novel deep active learning approach for object detection. Our approach relies on mixture density networks that estimate a probabilistic distribution for each localization and classification head's output. We explicitly estimate the aleatoric and epistemic uncertainty in a single forward pass of a single model. Our method

uses a scoring function that aggregates these two types of uncertainties for both heads to obtain every image's informativeness score. We demonstrate the efficacy of our approach in PASCAL VOC and MS-COCO datasets. Our approach outperforms single-model based methods and performs on par with multi-model based methods at a fraction of the computing cost.
****

Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID
Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11833-11842
Person Re-Identification (ReID) has achieved remarkable performance along with the deep learning era. However, most approaches carry out ReID only based upon holistic pedestrian regions. In contrast, real-world scenarios involve occluded pedestrians, which provide partial visual appearances and destroy the ReID accuracy. A common strategy is to locate visible body parts by auxiliary model, which however suffers from significant domain gaps and data bias issues. To avoid such problematic models in occluded person ReID, we propose the Occlusion-Aware Mask Network (OAMN). In particular, we incorporate an attention-guided mask module, which requires guidance from labeled occlusion data. To this end, we propose a novel occlusion augmentation scheme that produces diverse and precisely labeled occlusion for any holistic dataset. The proposed scheme suits real-world scenarios better than existing schemes, which consider only limited types of occlusions. We also offer a novel occlusion unification scheme to tackle ambiguity information at the test phase. The above three components enable existing attention mechanisms to precisely capture body parts regardless of the occlusion. Comprehensive experiments on a variety of person ReID benchmarks demonstrate the superiority of OAMN over state-of-the-arts.
****

Towards Accurate Alignment in Real-Time 3D Hand-Mesh Reconstruction
Xiao Tang, Tianyu Wang, Chi-Wing Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11698-11707
3D hand-mesh reconstruction from RGB images facilitates many applications, including augmented reality (AR). However, this requires not only real-time speed and accurate hand pose and shape but also plausible mesh-image alignment. While existing works already achieve promising results, meeting all three requirements is very challenging. This paper presents a novel pipeline by decoupling the hand-mesh reconstruction task into three stages: a joint stage to predict hand joints and segmentation; a mesh stage to predict a rough hand mesh; and a refine stage to fine-tune it with an offset mesh for mesh-image alignment. With careful design in the network structure and in the loss functions, we can promote high-quality finger-level mesh-image alignment and drive the models together to deliver real-time predictions. Extensive quantitative and qualitative results on benchmark datasets demonstrate that the quality of our results outperforms the state-of-the-art methods on hand-mesh/pose precision and hand-image alignment. In the end, we also showcase several real-time AR scenarios.
****

Searching for Controllable Image Restoration Networks
Heewon Kim, Sungyong Baik, Myungsub Choi, Janghoon Choi, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14234-14243
We present a novel framework for controllable image restoration that can effectively restore multiple types and levels of degradation of a corrupted image. The proposed model, named TASNet, is automatically determined by our neural architecture search algorithm, which optimizes the efficiency-accuracy trade-off of the candidate model architectures. Specifically, we allow TASNet to share the early layers across different restoration tasks and adaptively adjust the remaining layers with respect to each task. The shared task-agnostic layers greatly improve the efficiency while the task-specific layers are optimized for restoration quality, and our search algorithm seeks for the best balance between the two. We also propose a new data sampling strategy to further improve the overall restoratio

n performance. As a result, TASNet achieves significantly faster GPU latency and lower FLOPs compared to the existing state-of-the-art models, while also showing visually more pleasing outputs. The source code and pre-trained models are available at https://github.com/ghimhw/TASNet.

****

Cross-Category Video Highlight Detection via Set-Based Learning

Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, Changhu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7970-7979

Autonomous highlight detection is crucial for enhancing the efficiency of video browsing on social media platforms. To attain this goal in a data-driven way, one may often face the situation where highlight annotations are not available on the target video category used in practice, while the supervision on another video category (named as source video category) is achievable. In such a situation, one can derive an effective highlight detector on target video category by transferring the highlight knowledge acquired from source video category to the target one. We call this problem cross-category video highlight detection, which has been rarely studied in previous works. For tackling such practical problem, we propose a Dual-Learner-based Video Highlight Detection (DL-VHD) framework. Under this framework, we first design a Set-based Learning module (SL-module) to improve the conventional pair-based learning by assessing the highlight extent of a video segment under a broader context. Based on such learning manner, we introduce two different learners to acquire the basic distinction of target category videos and the characteristics of highlight moments on source video category, respectively. These two types of highlight knowledge are further consolidated via knowledge distillation. Extensive experiments on three benchmark datasets demonstrate the superiority of the proposed SL-module, and the DL-VHD method outperforms five typical Unsupervised Domain Adaptation (UDA) algorithms on various cross-category highlight detection tasks.

****

Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion

Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, Fengmao Lv; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8148-8156

Videos flow as the mixture of language, acoustic, and vision modalities. A thorough video understanding needs to fuse time-series data of different modalities for prediction. Due to the variable receiving frequency for sequences from each modality, there usually exists inherent asynchrony across the collected multimodal streams. Towards an efficient multimodal fusion from asynchronous multimodal streams, we need to model the correlations between elements from different modalities. The recent Multimodal Transformer (MulT) approach extends the self-attention mechanism of the original Transformer network to learn the crossmodal dependencies between elements. However, the direct replication of self-attention will suffer from the distribution mismatch across different modality features. As a result, the learnt crossmodal dependencies can be unreliable. Motivated by this observation, this work proposes the Modality-Invariant Crossmodal Attention (MICA) approach towards learning crossmodal interactions over modality-invariant space in which the distribution mismatch between different modalities is well bridged. To this end, both the marginal distribution and the elements with high-confidence correlations are aligned over the common space of the query and key vectors which are computed from different modalities. Experiments on three standard benchmarks of multimodal video understanding clearly validate the superiority of our approach.

****

Seeing Dynamic Scene in the Dark: A High-Quality Video Dataset With Mechatronic Alignment

Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9700-9709

Low-light video enhancement is an important task. Previous work is mostly traine

d on paired static images or videos. We compile a new dataset formed by our new strategy that contains high-quality spatially-aligned video pairs from dynamic s cenes in low- and normal-light conditions. We built it using a mechatronic syste m to precisely control the dynamics during the video capture process, and furthe r align the video pairs, both spatially and temporally, by identifying the syste m's uniform motion stage. Besides the dataset, we propose an end-to-end framewor k, in which we design a self-supervised strategy to reduce noise, while enhancin g the illumination based on the Retinex theory. Extensive experiments based on v arious metrics and large-scale user study demonstrate the value of our dataset a nd effectiveness of our method. The dataset and code are available at https://gi thub.com/dvlab-research/SDSD.
****

AdvRush: Searching for Adversarially Robust Neural Architectures
Jisoo Mok, Byunggook Na, Hyeokjun Choe, Sungroh Yoon; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 12322-12332
Deep neural networks continue to awe the world with their remarkable performance . Their predictions, however, are prone to be corrupted by adversarial examples that are imperceptible to humans. Current efforts to improve the robustness of n eural networks against adversarial examples are focused on developing robust tra ining methods, which update the weights of a neural network in a more robust dir ection. In this work, we take a step beyond training of the weight parameters an d consider the problem of designing an adversarially robust neural architecture with high intrinsic robustness. We propose AdvRush, a novel adversarial robustne ss-aware neural architecture search algorithm, based upon a finding that indepen dent of the training method, the intrinsic robustness of a neural network can be represented with the smoothness of its input loss landscape. Through a regulari zer that favors a candidate architecture with a smoother input loss landscape, A dvRush successfully discovers an adversarially robust neural architecture. Along with a comprehensive theoretical motivation for AdvRush, we conduct an extensiv e amount of experiments to demonstrate the efficacy of AdvRush on various benchm ark datasets. Notably, on CIFAR-10, AdvRush achieves 55.91% robust accuracy unde r FGSM attack after standard training and 50.04% robust accuracy under AutoAttac k after 7-step PGD adversarial training.
****

Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Net works in Frequency Domain
Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 458-4 67
Recently, the generalization behavior of Convolutional Neural Networks (CNN) is gradually transparent through explanation techniques with the frequency componen ts decomposition. However, the importance of the phase spectrum of the image for a robust vision system is still ignored. In this paper, we notice that the CNN tends to converge at the local optimum which is closely related to the high-freq uency components of the training images, while the amplitude spectrum is easily disturbed such as noises or common corruptions. In contrast, more empirical stud ies found that humans rely on more phase components to achieve robust recognitio n. This observation leads to more explanations of the CNN's generalization behav iors in both robustness to common perturbations and out-of-distribution detectio n, and motivates a new perspective on data augmentation designed by re-combing t he phase spectrum of the current image and the amplitude spectrum of the distrac ter image. That is, the generated samples force the CNN to pay more attention to the structured information from phase components and keep robust to the variati on of the amplitude. Experiments on several image datasets indicate that the pro posed method achieves state-of-the-art performances on multiple generalizations and calibration tasks, including adaptability for common corruptions and surface variations, out-of-distribution detection, and adversarial attack.
****

Mean Shift for Self-Supervised Learning
Soroush Abbasi Koohpayegani, Ajinkya Tejankar, Hamed Pirsiavash; Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10326-10335
Most recent self-supervised learning (SSL) algorithms learn features by contrasting between instances of images or by clustering the images and then contrasting between the image clusters. We introduce a simple mean-shift algorithm that learns representations by grouping images together without contrasting between them or adopting much of prior on the structure or number of the clusters. We simply "shift" the embedding of each image to be close to the "mean" of the neighbors of its augmentation. Since the closest neighbor is always another augmentation of the same image, our model will be identical to BYOL when using only one nearest neighbor instead of 5 used in our experiments. Our model achieves 72.4% on ImageNet linear evaluation with ResNet50 at 200 epochs outperforming BYOL. Also, our method outperforms the SOTA by a large margin when using weak augmentations only, facilitating the adoption of SSL for other modalities. Our code is available here: https://github.com/UMBCvision/MSF
****

Speech Drives Templates: Co-Speech Gesture Synthesis With Learned Templates
Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, Shenghua Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11077-11086
Co-speech gesture generation is to synthesize a gesture sequence that not only looks real but also matches with the input speech audio. Our method generates the movements of a complete upper body, including arms, hands, and the head. Although recent data-driven methods achieve great success, challenges still exist, such as limited variety, poor fidelity, and lack of objective metrics. Motivated by the fact that the speech cannot fully determine the gesture, we design a method that learns a set of gesture template vectors to model the latent conditions, which relieve the ambiguity. For our method, the template vector determines the general appearance of a generated gesture sequence, while the speech audio drives subtle movements of the body, both indispensable for synthesizing a realistic gesture sequence. Due to the intractability of an objective metric for gesture-speech synchronization, we adopt the lip-sync error as a proxy metric to tune and evaluate the synchronization ability of our model. Extensive experiments show the superiority of our method in both objective and subjective evaluations on fidelity and synchronization.
****

Improving Robustness Against Common Corruptions With Frequency Biased Models
Tonmoy Saikia, Cordelia Schmid, Thomas Brox; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10211-10220
CNNs perform remarkably well when the training and test distributions are i.i.d, but unseen image corruptions can cause a surprisingly large drop in performance. In various real scenarios, unexpected distortions, such as random noise, compression artefacts, or weather distortions are common phenomena. Improving performance on corrupted images must not result in degraded i.i.d performance - a challenge faced by many state-of-the-art robust approaches. Image corruption types have different characteristics in the frequency spectrum and would benefit from a targeted type of data augmentation, which, however, is often unknown during training. In this paper, we introduce a mixture of two expert models specializing in high and low-frequency robustness, respectively. Moreover, we propose a new regularization scheme that minimizes the total variation (TV) of convolution feature-maps to increase high-frequency robustness. The approach improves on corrupted images without degrading in-distribution performance. We demonstrate this on ImageNet-C and also for real-world corruptions on an automotive dataset, both for object classification and object detection.
****

AdvDrop: Adversarial Attack to DNNs by Dropping Information
Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A. K. Qin, Yuan He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7506-7515
Human can easily recognize visual objects with lost information: even losing most details with only contour reserved, e.g. cartoon. However, in terms of visual

perception of Deep Neural Networks (DNNs), the ability for recognizing abstract objects (visual objects with lost information) is still a challenge. In this work, we investigate this issue from an adversarial viewpoint: will the performance of DNNs decrease even for the images only losing a little information? Towards this end, we propose a novel adversarial attack, named AdvDrop, which crafts adversarial examples by dropping existing information of images. Previously, most adversarial attacks add extra disturbing information on clean images explicitly. Opposite to previous works, our proposed work explores the adversarial robustness of DNN models in a novel perspective by dropping imperceptible details to craft adversarial examples. We demonstrate the effectiveness of AdvDrop by extensive experiments, and show that this new type of adversarial examples is more difficult to be defended by current defense systems.

****

HuMoR: 3D Human Motion Model for Robust Pose Estimation

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11488-11499

We introduce HuMoR: a 3D Human Motion Model for Robust Estimation of temporal pose and shape. Though substantial progress has been made in estimating 3D human motion and shape from dynamic observations, recovering plausible pose sequences in the presence of noise and occlusions remains a challenge. For this purpose, we propose an expressive generative model in the form of a conditional variational autoencoder, which learns a distribution of the change in pose at each step of a motion sequence. Furthermore, we introduce a flexible optimization-based approach that leverages HuMoR as a motion prior to robustly estimate plausible pose and shape from ambiguous observations. Through extensive evaluations, we demonstrate that our model generalizes to diverse motions and body shapes after training on a large motion capture dataset, and enables motion reconstruction from multiple input modalities including 3D keypoints and RGB(-D) videos.

****

Class Semantics-Based Attention for Action Detection

Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, Juwei Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13739-13748

Action localization networks are often structured as a feature encoder sub-network and a localization sub-network, where the feature encoder learns to transform an input video to features that are useful for the localization sub-network to generate reliable action proposals. While some of the encoded features may be more useful for generating action proposals, prior action localization approaches do not include any attention mechanism that enables the localization sub-network to attend more to the more important features. In this paper, we propose a novel attention mechanism, the Class Semantics-based Attention (CSA), that learns from the temporal distribution of semantics of action classes present in an input video to find the importance scores of the encoded features, which are used to provide attention to the more useful encoded features. We demonstrate on two popular action detection datasets that incorporating our novel attention mechanism provides considerable performance gains on competitive action detection models (e.g., around 6.2% improvement over BMN action detection baseline to obtain 47.5% mAP on the THUMOS-14 dataset), and a new state-of-the-art of 36.25% mAP on the ActivityNet v1.3 dataset. Further, the CSA localization model family which includes BMN-CSA, was part of the second-placed submission at the 2021 ActivityNet action localization challenge. Our attention mechanism outperforms prior self-attention modules such as the squeeze-and-excitation in action detection task. We also observe that our attention mechanism is complementary to such self-attention modules in that performance improvements are seen when both are used together.

****

Adaptive Graph Convolution for Point Cloud Analysis

Haoran Zhou, Yidan Feng, Mingsheng Fang, Mingqiang Wei, Jing Qin, Tong Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4965-4974

Convolution on 3D point clouds that generalized from 2D grid-like domains is widely researched yet far from perfect. The standard convolution characterises feature correspondences indistinguishably among 3D points, presenting an intrinsic limitation of poor distinctive feature learning. In this paper, we propose Adaptive Graph Convolution (AdaptConv) which generates adaptive kernels for points according to their dynamically learned features. Compared with using a fixed/isotropic kernel, AdaptConv improves the flexibility of point cloud convolutions, effectively and precisely capturing the diverse relations between points from different semantic parts. Unlike popular attentional weight schemes, the proposed AdaptConv implements the adaptiveness inside the convolution operation instead of simply assigning different weights to the neighboring points. Extensive qualitative and quantitative evaluations show that our method outperforms state-of-the-art point cloud classification and segmentation approaches on several benchmark datasets. Our code is available at https://github.com/hrzhou2/AdaptConv-master.

****

Adversarial Attack on Deep Cross-Modal Hamming Retrieval

Chao Li, Shangqian Gao, Cheng Deng, Wei Liu, Heng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2218-2227

Recently, Cross-Modal Hamming space Retrieval (CMHR) regains ever-increasing attention, mainly benefiting from the excellent representation capability of deep neural networks. On the other hand, the vulnerability of deep networks exposes a deep cross-modal retrieval system to various safety risks (e.g., adversarial attack). However, attacking deep cross-modal Hamming retrieval remains underexplored. In this paper, we propose an effective Adversarial Attack on Deep Cross-Modal Hamming Retrieval, dubbed AACH, which fools a target deep CMHR model in a black-box setting. Specifically, given a target model, we first construct its substitute model to exploit cross-modal correlations within hamming space, with which we create adversarial examples by limitedly querying from a target model. Furthermore, to enhance the efficiency of adversarial attacks, we design a triplet construction module to exploit cross-modal positive and negative instances. In this way, perturbations can be learned to fool the target model through pulling perturbed examples far away from the positive instances whereas pushing them close to the negative ones. Extensive experiments on three widely used cross-modal (image and text) retrieval benchmarks demonstrate the superiority of the proposed AACH. We find that AACH can successfully attack a given target deep CMHR model with fewer interactions, and that its performance is on par with previous state-of-the-art attacks.

****

UASNet: Uncertainty Adaptive Sampling Network for Deep Stereo Matching

Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, Hong-Seok Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6311-6319

Recent studies have shown that cascade cost volume can play a vital role in deep stereo matching to achieve high resolution depth map with efficient hardware usage. However, how to construct good cascade volume as well as effective sampling for them are still under in-depth study. Previous cascade-based methods usually perform uniform sampling in a predicted disparity range based on variance, which easily misses the ground truth disparity and decreases disparity map accuracy. In this paper, we propose an uncertainty adaptive sampling network (UASNet) featuring two modules: an uncertainty distribution-guided range prediction (URP) model and an uncertainty-based disparity sampler (UDS) module. The URP explores the more discriminative uncertainty distribution to handle the complex matching ambiguities and to improve disparity range prediction. The UDS adaptively adjusts sampling interval to localize disparity with improved accuracy. With the proposed modules, our UASNet learns to construct cascade cost volume and predict full-resolution disparity map directly. Extensive experiments show that the proposed method achieves the highest ground truth covering ratio compared with other cascade cost volume based stereo matching methods. Our method also achieves top performance on both SceneFlow dataset and KITTI benchmark.

****

Minimal Adversarial Examples for Deep Learning on 3D Point Clouds
Jaeyeon Kim, Binh-Son Hua, Thanh Nguyen, Sai-Kit Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7797-7806
With recent developments of convolutional neural networks, deep learning for 3D point clouds has shown significant progress in various 3D scene understanding tasks, e.g., object recognition, object detection. In a safety-critical environment, it is however not well understood how such deep learning models are vulnerable to adversarial examples. In this work, we explore adversarial attacks for point cloud-based neural networks. We propose a new formulation for adversarial point cloud generation that can generalise two different attack strategies. Our method generates adversarial examples by attacking the classification ability of point cloud-based networks while considering the perceptibility of the examples and ensuring the minimal level of point manipulations. Experimental results show that our method achieves the state-of-the-art performance with higher than 89% and 90% of attack success rate on synthetic and real-world data respectively, while manipulating only about 4% of the total points.
****
MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions
Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13536-13545
Spatio-temporal action detection is an important and challenging problem in video understanding. The existing action detection benchmarks are limited in aspects of small numbers of instances in a trimmed video or low-level atomic actions. This paper aims to present a new multi-person dataset of spatio-temporal localized sports actions, coined as MultiSports. We first analyze the important ingredients of constructing a realistic and challenging dataset for spatio-temporal action detection by proposing three criteria: (1) multi-person scenes and motion dependent identification, (2) with well-defined boundaries, (3) relatively fine-grained classes of high complexity. Based on these guidelines, we build the dataset of MultiSports v1.0 by selecting 4 sports classes, collecting 3200 video clips, and annotating 37701 action instances with 902k bounding boxes. Our datasets are characterized with important properties of high diversity, dense annotation, and high quality. Our MultiSports, with its realistic setting and detailed annotations, exposes the intrinsic challenges of spatio-temporal action detection. To benchmark this, we adapt several baseline methods to our dataset and give an in-depth analysis on the action detection results in our dataset. We hope our MultiSports can serve as a standard benchmark for spatio-temporal action detection in the future. Our dataset website is at https://deeperaction.github.io/multisports/.
****
Triggering Failures: Out-of-Distribution Detection by Learning From Local Adversarial Attacks in Semantic Segmentation
Victor Besnier, Andrei Bursuc, David Picard, Alexandre Briot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15701-15710
In this paper, we tackle the detection of out-of-distribution (OOD) objects in semantic segmentation. By analyzing the literature, we found that current methods are either accurate or fast but not both which limits their usability in real world applications. To get the best of both aspects, we propose to mitigate the common shortcomings by following four design principles: decoupling the OOD detection from the segmentation task, observing the entire segmentation network instead of just its output, generating training data for the OOD detector by leveraging blind spots in the segmentation network and focusing the generated data on localized regions in the image to simulate OOD objects. Our main contribution is a new OOD detection architecture called ObsNet associated with a dedicated training scheme based on Local Adversarial Attacks (LAA). We validate the soundness of our approach across numerous ablation studies. We also show it obtains top performances both in speed and accuracy when compared to ten recent methods of the l

iterature on three different datasets.
****
Glimpse-Attend-and-Explore: Self-Attention for Active Visual Exploration
Soroush Seifi, Abhishek Jha, Tinne Tuytelaars; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 16137-16146
Active visual exploration aims to assist an agent with a limited field of view t
o understand its environment based on partial observations made by choosing the
best viewing directions in the scene. Recent methods have tried to address this
problem either by using reinforcement learning, which is difficult to train, or
by uncertainty maps, which are task-specific and can only be implemented for den
se prediction tasks. In this paper, we propose the Glimpse-Attend-and-Explore mo
del which: (a) employs self-attention to guide the visual exploration instead of
 task-specific uncertainty maps; (b) can be used for both dense and sparse predi
ction tasks; and (c) uses a contrastive stream to further improve the representa
tions learned. Unlike previous works, we show the application of our model on mu
ltiple tasks like reconstruction, segmentation and classification. Our model pro
vides encouraging results against baseline while being less dependent on dataset
 bias in driving the exploration. We further perform an ablation study to invest
igate the features and attention learned by our model. Finally, we show that our
 self-attention module learns to attend different regions of the scene by minimi
zing the loss on the downstream task. Code: https://github.com/soroushseifi/glim
pse-attend-explore.
****
Field Convolutions for Surface CNNs
Thomas W. Mitchel, Vladimir G. Kim, Michael Kazhdan; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 10001-10011
We present a novel surface convolution operator acting on vector fields that is
based on a simple observation: instead of combining neighboring features with re
spect to a single coordinate parameterization defined at a given point, we have
every neighbor describe the position of the point within its own coordinate fram
e. This formulation combines intrinsic spatial convolution with parallel transpo
rt in a scattering operation while placing no constraints on the filters themsel
ves, providing a definition of convolution that commutes with the action of isom
etries, has increased descriptive potential, and is robust to noise and other nu
isance factors. The result is a rich notion of convolution which we call field c
onvolution, well-suited for CNNs on surfaces. Field convolutions are flexible, s
traight-forward to incorporate into surface learning frameworks, and their highl
y discriminating nature has cascading effects throughout the learning pipeline.
Using simple networks constructed from residual field convolution blocks, we ach
ieve state-of-the-art results on standard benchmarks in fundamental geometry pro
cessing tasks, such as shape classification, segmentation, correspondence, and s
parse matching.
****
SIMstack: A Generative Shape and Instance Model for Unordered Object Stacks
Zoe Landgraf, Raluca Scona, Tristan Laidlow, Stephen James, Stefan Leutenegger,
Andrew J. Davison; Proceedings of the IEEE/CVF International Conference on Compu
ter Vision (ICCV), 2021, pp. 13012-13022
By estimating 3D shape and instances from a single view, we can capture informat
ion about the environment quickly, without the need for comprehensive scanning a
nd multi-view fusion. Solving this task for composite scenes (such as object sta
cks) is challenging: occluded areas are not only ambiguous in shape but also in
instance segmentation; multiple decompositions could be valid. We observe that p
hysics constrains decomposition as well as shape in occluded regions and hypothe
sise that a latent space learned from scenes built under physics simulation can
serve as a prior to better predict shape and instances in occluded regions. To t
his end we propose SIMstack, a depth-conditioned Variational Auto-Encoder (VAE),
 trained on a dataset of objects stacked under physics simulation. We formulate
instance segmentation as a center voting task which allows for class-agnostic de
tection and doesn't require setting the maximum number of objects in the scene.
At test time, our model can generate 3D shape and instance segmentation from a s

ingle depth view, probabilistically sampling proposals for the occluded region f rom the learned latent space. We argue that this method has practical applicatio ns in providing robots some of the ability humans have to make rapid intuitive i nferences of partially observed scenes. We demonstrate an application for precis e (non-disruptive) object grasping of unknown objects from a single depth view.
****

Weakly Supervised Person Search With Region Siamese Networks
Chuchu Han, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang, Changhu Wang; Proceedings of the IEEE/CVF International Conference on Computer V ision (ICCV), 2021, pp. 12006-12015
Supervised learning is dominant in person search, but it requires elaborate labe ling of bounding boxes and identities. Large-scale labeled training data is ofte n difficult to collect, especially for person identities. A natural question is whether a good person search model can be trained without the need of identity s upervision. In this paper, we present a weakly supervised setting where only bou nding box annotations are available. Based on this new setting, we provide an ef fective baseline model termed Region Siamese Networks (R-SiamNets). Towards lear ning useful representations for recognition in the absence of identity labels, w e supervise the R-SiamNet with instance-level consistency loss and cluster-level contrastive loss. For instance-level consistency learning, the R-SiamNet is con strained to extract consistent features from each person region with or without out-of-region context. For cluster-level contrastive learning, we enforce the ag gregation of closest instances and the separation of dissimilar ones in feature space. Extensive experiments validate the utility of our weakly supervised metho d. Our model achieves the rank-1 of 87.1% and mAP of 86.0% on CUHK-SYSU benchmar k, which surpasses several fully supervised methods, such as OIM and MGTS, by a clear margin. More promising performance can be reached by incorporating extra t raining data. We hope this work could encourage the future research in this fiel d.
****

Learning Icosahedral Spherical Probability Map Based on Bingham Mixture Model fo r Vanishing Point Estimation
Haoang Li, Kai Chen, Pyojin Kim, Kuk-Jin Yoon, Zhe Liu, Kyungdon Joo, Yun-Hui Li u; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5661-5670
Existing vanishing point (VP) estimation methods rely on pre-extracted image lin es and/or prior knowledge of the number of VPs. However, in practice, this infor mation may be insufficient or unavailable. To solve this problem, we propose a n etwork that treats a perspective image as input and predicts a spherical probabi lity map of VP. Based on this map, we can detect all the VPs. Our method is reli able thanks to four technical novelties. First, we leverage the icosahedral sphe rical representation to express our probability map. This representation provide s uniform pixel distribution, and thus facilitates estimating arbitrary position s of VPs. Second, we design a loss function that enforces the antipodal symmetry and sparsity of our spherical probability map to prevent over-fitting. Third, w e generate the ground truth probability map that reasonably expresses the locati ons and uncertainties of VPs. This map unnecessarily peaks at noisy annotated VP s, and also exhibits various anisotropic dispersions. Fourth, given a predicted probability map, we detect VPs by fitting a Bingham mixture model. This strategy can robustly handle close VPs and provide the confidence level of VP useful for practical applications. Experiments showed that our method achieves the best co mpromise between generality, accuracy, and efficiency, compared with state-of-th e-art approaches.
****

Incorporating Learnable Membrane Time Constant To Enhance Learning of Spiking Ne ural Networks
Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, Yonghong Ti an; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2021, pp. 2661-2671
Spiking Neural Networks (SNNs) have attracted enormous research interest due to

temporal information processing capability, low power consumption, and high biological plausibility. However, the formulation of efficient and high-performance learning algorithms for SNNs is still challenging. Most existing learning methods learn weights only, and require manual tuning of the membrane-related parameters that determine the dynamics of a single spiking neuron. These parameters are typically chosen to be the same for all neurons, which limits the diversity of neurons and thus the expressiveness of the resulting SNNs. In this paper, we take inspiration from the observation that membrane-related parameters are different across brain regions, and propose a training algorithm that is capable of learning not only the synaptic weights but also the membrane time constants of SNNs. We show that incorporating learnable membrane time constants can make the network less sensitive to initial values and can speed up learning. In addition, we reevaluate the pooling methods in SNNs and find that max-pooling will not lead to significant information loss and have the advantage of low computation cost and binary compatibility. We evaluate the proposed method for image classification tasks on both traditional static MNIST, Fashion-MNIST, CIFAR-10 datasets, and neuromorphic N-MNIST, CIFAR10-DVS, DVS128 Gesture datasets. The experiment results show that the proposed method outperforms the state-of-the-art accuracy on nearly all datasets, using fewer time-steps. Our codes are available at https://github.com/fangwei123456/Parametric-Leaky-Integrate-and-Fire-Spiking-Neuron.

****

Real-Time Vanishing Point Detector Integrating Under-Parameterized RANSAC and Hough Transform

Jianping Wu, Liang Zhang, Ye Liu, Ke Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3732-3741

We propose a novel approach that integrates under-parameterized RANSAC (UPRANSAC) with Hough Transform to detect vanishing points (VPs) from un-calibrated monocular images. In our algorithm, the UPRANSAC chooses one hypothetical inlier in a sample set to find a portion of the VP's degrees of freedom, which is followed by a highly reliable brute-force voting scheme (1-D Hough Transform) to find the VP's remaining degrees of freedom along the extension line of the hypothetical inlier. Our approach is able to sequentially find a series of VPs by repeatedly removing inliers of any detected VPs from minimal sample sets until the stop criterion is reached. Compared to traditional RANSAC that selects 2 edges as a hypothetical inlier pair to fit a model of VP hypothesis and requires hitting a pair of inliners, the UPRANSAC has a higher likelihood to hit one inliner and is more reliable in VP detection. Meanwhile, the tremendously scaled-down voting space with the requirement of only 1 parameter for processing significantly increased the performance efficiency of Hough Transform in our scheme. Testing results with well-known benchmark datasets show that the detection accuracies of our approach were higher or on par with the SOTA while running in deeply real-time zone.

****

Pose Invariant Topological Memory for Visual Navigation

Asuto Taniguchi, Fumihiro Sasaki, Ryota Yamashina; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15384-15393

Planning for visual navigation using topological memory, a memory graph consisting of nodes and edges, has been recently well-studied. The nodes correspond to past observations of a robot, and the edges represent the reachability predicted by a neural network (NN). Most prior methods, however, often fail to predict the reachability when the robot takes different poses, i.e. the direction the robot faces, at close positions. This is because the methods observe first-person view images, which significantly changes when the robot changes its pose, and thus it is fundamentally difficult to correctly predict the reachability from them. In this paper, we propose pose invariant topological memory (POINT) to address the problem. POINT observes omnidirectional images and predicts the reachability by using a spherical convolutional NN, which has a rotation invariance property and enables planning regardless of the robot's pose. Additionally, we train the NN by contrastive learning with data augmentation to enable POINT to plan with robustness to changes in environmental conditions, such as light conditions and the presence of unseen objects. Our experimental results show that POINT outperfor

ms conventional methods under both the same and different environmental conditio
ns. In addition, the results with the KITTI-360 dataset show that POINT is more
applicable to real-world environments than conventional methods.
****

Shape Self-Correction for Unsupervised Point Cloud Understanding
Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li,
 Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2021, pp. 8382-8391
We develop a novel self-supervised learning method named Shape Self-Correction f
or point cloud analysis. Our method is motivated by the principle that a good sh
ape representation should be able to find distorted parts of a shape and correct
 them. To learn strong shape representations in an unsupervised manner, we first
 design a shape-disorganizing module to destroy certain local shape parts of an
object. Then the destroyed shape and the normal shape are sent into a point clou
d network to get representations, which are employed to segment points that belo
ng to distorted parts and further reconstruct them to restore the shape to norma
l. To perform better in these two associated pretext tasks, the network is const
rained to capture useful shape features from the object, which indicates that th
e point cloud network encodes rich geometric and contextual information. The lea
rned feature extractor transfers well to downstream classification and segmentat
ion tasks. Experimental results on ModelNet, ScanNet and ShapeNetPart demonstrat
e that our method achieves state-of-the-art performance among unsupervised metho
ds. Our framework can be applied to a wide range of deep learning networks for p
oint cloud analysis and we show experimentally that pre-training with our framew
ork significantly boosts the performance of supervised models.
****

ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement
Yuval Alaluf, Or Patashnik, Daniel Cohen-Or; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2021, pp. 6711-6720
Recently, the power of unconditional image synthesis has significantly advanced
through the use of Generative Adversarial Networks (GANs). The task of inverting
 an image into its corresponding latent code of the trained GAN is of utmost imp
ortance as it allows for the manipulation of real images, leveraging the rich se
mantics learned by the network. Recognizing the limitations of current inversion
 approaches, in this work we present a novel inversion scheme that extends curre
nt encoder-based inversion methods by introducing an iterative refinement mechan
ism. Instead of directly predicting the latent code of a given real image using
a single pass, the encoder is tasked with predicting a residual with respect to
the current estimate of the inverted latent code in a self-correcting manner. Ou
r residual-based encoder, named ReStyle, attains improved accuracy compared to c
urrent state-of-the-art encoder-based methods with a negligible increase in infe
rence time. We analyze the behavior of ReStyle to gain valuable insights into it
s iterative nature. We then evaluate the performance of our residual encoder and
 analyze its robustness compared to optimization-based inversion and state-of-th
e-art encoders.
****

Low-Rank Tensor Completion by Approximating the Tensor Average Rank
Zhanliang Wang, Junyu Dong, Xinguo Liu, Xueying Zeng; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2021, pp. 4612-4620
This paper focuses on the problem of low-rank tensor completion, the goal of whi
ch is to recover an underlying low-rank tensor from incomplete observations. Our
 method is motivated by the recently proposed t-product based on any invertible
linear transforms. First, we define the new tensor average rank under the invert
ible real linear transforms. We then propose a new tensor completion model using
 a nonconvex surrogate to approximate the tensor average rank. This surrogate ov
ercomes the discontinuity of the tensor average rank and alleviates the bias pro
blem caused by the convex relaxation. Further, we develop an efficient algorithm
 to solve the proposed model and establish its convergence. Finally, experimenta
l results on both synthetic and real data demonstrate the superiority of our met
hod.

****
Dissecting Image Crops
Basile Van Hoorick, Carl Vondrick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9741-9750
The elementary operation of cropping underpins nearly every computer vision system, ranging from data augmentation and translation invariance to computational photography and representation learning. This paper investigates the subtle traces introduced by this operation. For example, despite refinements to camera optics, lenses will leave behind certain clues, notably chromatic aberration and vignetting. Photographers also leave behind other clues relating to image aesthetics and scene composition. We study how to detect these traces, and investigate the impact that cropping has on the image distribution. While our aim is to dissect the fundamental impact of spatial crops, there are also a number of practical implications to our work, such as revealing faulty photojournalism and equipping neural network researchers with a better understanding of shortcut learning. Code is available at https://github.com/basilevh/dissecting-image-crops.
****
Exploiting Multi-Object Relationships for Detecting Adversarial Attacks in Complex Scenes
Mingjun Yin, Shasha Li, Zikui Cai, Chengyu Song, M. Salman Asif, Amit K. Roy-Chowdhury, Srikanth V. Krishnamurthy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7858-7867
Vision systems that deploy Deep Neural Networks (DNNs) are known to be vulnerable to adversarial examples. Recent research has shown that checking the intrinsic consistencies in the input data is a promising way to detect adversarial attacks (e.g., by checking the object co-occurrence relationships in complex scenes). However, existing approaches are tied to specific models and do not offer generalizability. Motivated by the observation that language descriptions of natural scene images have already captured the object co-occurrence relationships that can be learned by a language model, we develop a novel approach to perform context consistency checks using such language models. The distinguishing aspect of our approach is that it is independent of the deployed object detector and yet offers very high accuracy in terms of detecting adversarial examples in practical scenes with multiple objects. Experiments on the PASCAL VOC and MS COCO datasets show that our method can outperform state-of-the-art methods in detecting adversarial attacks.
****
Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation
Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, Yu-Xiong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7273-7282
We present a novel semi-supervised semantic segmentation method which jointly achieves two desiderata of segmentation model regularities: the label-space consistency property between image augmentations and the feature-space contrastive property among different pixels. We leverage the pixel-level L2 loss and the pixel contrastive loss for the two purposes respectively. To address the computational efficiency issue and the false negative noise issue involved in the pixel contrastive loss, we further introduce and investigate several negative sampling techniques. Extensive experiments demonstrate the state-of-the-art performance of our method (PC2Seg) with the DeepLab-v3+ architecture, in several challenging semi-supervised settings derived from the VOC, Cityscapes, and COCO datasets.
****
Standardized Max Logits: A Simple yet Effective Approach for Identifying Unexpected Road Obstacles in Urban-Scene Segmentation
Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15425-15434
Identifying unexpected objects on roads in semantic segmentation (e.g., identifying dogs on roads) is crucial in safety-critical applications. Existing approaches use images of unexpected objects from external datasets or require additional

training (e.g., retraining segmentation networks or training an extra network), which necessitate a non-trivial amount of labor intensity or lengthy inference time. One possible alternative is to use prediction scores of a pre-trained network such as the max logits (i.e., maximum values among classes before the final softmax layer) for detecting such objects. However, the distribution of max logits of each predicted class is significantly different from each other, which degrades the performance of identifying unexpected objects in urban-scene segmentation. To address this issue, we propose a simple yet effective approach that standardizes the max logits in order to align the different distributions and reflect the relative meanings of max logits within each predicted class. Moreover, we consider the local regions from two different perspectives based on the intuition that neighboring pixels share similar semantic information. In contrast to previous approaches, our method does not utilize any external datasets or require additional training, which makes our method widely applicable to existing pre-trained segmentation models. Such a straightforward approach achieves a new state-of-the-art performance on the publicly available Fishyscapes Lost & Found leaderboard with a large margin. Our code is publicly available at this link.

****

SIGNET: Efficient Neural Representation for Light Fields

Brandon Yushan Feng, Amitabh Varshney; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14224-14233

We present a novel neural representation for light field content that enables compact storage and easy local reconstruction with high fidelity. We use a fully-connected neural network to learn the mapping function between each light field pixel's coordinates and its corresponding color values. However, neural networks that simply take in raw coordinates are unable to accurately learn data containing fine details. We present an input transformation strategy based on the Gegenbauer polynomials which previously showed theoretical advantages over the Fourier basis. We conduct experiments that show our Gegenbauer-based design combined with sinusoidal activation functions leads to a better light field reconstruction quality than a variety of network designs, including those with Fourier-inspired techniques introduced by prior works. Moreover, our SInusoidal Gegenbauer NETwork, or SIGNET, can represent light field scenes more compactly than the state-of-the-art compression methods while maintaining a comparable reconstruction quality. SIGNET also innately allows random access to encoded light field pixels due to its functional design. Furthermore, we demonstrate that SIGNET facilitates super-resolution along the spatial, angular, and temporal dimensions of a light field without any additional training.

****

Cross-Descriptor Visual Localization and Mapping

Mihai Dusmanu, Ondrej Miksik, Johannes L. Schönberger, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6058-6067

Visual localization and mapping is the key technology underlying the majority of mixed reality and robotics systems. Most state-of-the-art approaches rely on local features to establish correspondences between images. In this paper, we present three novel scenarios for localization and mapping which require the continuous update of feature representations and the ability to match across different feature types. While localization and mapping is a fundamental computer vision problem, the traditional setup supposes the same local features are used throughout the evolution of a map. Thus, whenever the underlying features are changed, the whole process is repeated from scratch. However, this is typically impossible in practice, because raw images are often not stored and re-building the maps could lead to loss of the attached digital content. To overcome the limitations of current approaches, we present the first principled solution to cross-descriptor localization and mapping. Our data-driven approach is agnostic to the feature descriptor type, has low computational requirements, and scales linearly with the number of description algorithms. Extensive experiments demonstrate the effectiveness of our approach on state-of-the-art benchmarks for a variety of handcrafted and learned features.

****
Understanding and Evaluating Racial Biases in Image Captioning
Dora Zhao, Angelina Wang, Olga Russakovsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14830-14840
Image captioning is an important task for benchmarking visual reasoning and for enabling accessibility for people with vision impairments. However, as in many machine learning settings, social biases can influence image captioning in undesirable ways. In this work, we study bias propagation pathways within image captioning, focusing specifically on the COCO dataset. Prior work has analyzed gender bias in captions using automatically-derived gender labels; here we examine racial and intersectional biases using manual annotations. Our first contribution is in annotating the perceived gender and skin color of 28,315 of the depicted people after obtaining IRB approval. Using these annotations, we compare racial biases present in both manual and automatically-generated image captions. We demonstrate differences in caption performance, sentiment, and word choice between images of lighter versus darker-skinned people. Further, we find the magnitude of these differences to be greater in modern captioning systems compared to older ones, thus leading to concerns that without proper consideration and mitigation these differences will only become increasingly prevalent. Code and data is available at https://princetonvisualai.github.io/imagecaptioning-bias/.
****
Panoptic Narrative Grounding
Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, Pablo Arbeláez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1364-1373
This paper proposes Panoptic Narrative Grounding, a spatially fine and general formulation of the natural language visual grounding problem. We establish an experimental framework for the study of this new task, including new ground truth and metrics, and we propose a strong baseline method to serve as stepping stone for future work. We exploit the intrinsic semantic richness in an image by including panoptic categories, and we approach visual grounding at a fine-grained level by using segmentations. In terms of ground truth, we propose an algorithm to automatically transfer Localized Narratives annotations to specific regions in the panoptic segmentations of the MS COCO dataset. To guarantee the quality of our annotations, we take advantage of the semantic structure contained in WordNet to exclusively incorporate noun phrases that are grounded to a meaningfully related panoptic segmentation region. The proposed baseline achieves a performance of 55.4 absolute Average Recall points. This result is a suitable foundation to push the envelope further in the development of methods for Panoptic Narrative Grounding.
****
Weakly-Supervised Video Anomaly Detection With Robust Temporal Feature Magnitude Learning
Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, Gustavo Carneiro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4975-4986
Anomaly detection with weakly supervised video-level labels is typically formulated as a multiple instance learning (MIL) problem, in which we aim to identify snippets containing abnormal events, with each video represented as a bag of video snippets. Although current methods show effective detection performance, their recognition of the positive instances, i.e., rare abnormal snippets in the abnormal videos, is largely biased by the dominant negative instances, especially when the abnormal events are subtle anomalies that exhibit only small differences compared with normal events. This issue is exacerbated in many methods that ignore important video temporal dependencies. To address this issue, we introduce a novel and theoretically sound method, named Robust Temporal Feature Magnitude learning (RTFM), which trains a feature magnitude learning function to effectively recognise the positive instances, substantially improving the robustness of the MIL approach to the negative instances from abnormal videos. RTFM also adapts dilated convolutions and self-attention mechanisms to capture long- and short-ran

ge temporal dependencies to learn the feature magnitude more faithfully. Extensive experiments show that the RTFM-enabled MIL model (i) outperforms several state-of-the-art methods by a large margin on four benchmark data sets (ShanghaiTech, UCF-Crime, XD-Violence and UCSD-Peds) and (ii) achieves significantly improved subtle anomaly discriminability and sample efficiency.
****

Learning an Augmented RGB Representation With Cross-Modal Knowledge Distillation for Action Detection
Rui Dai, Srijan Das, François Bremond; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13053-13064
In video understanding, most cross-modal knowledge distillation (KD) methods are tailored for classification tasks, focusing on the discriminative representation of the trimmed videos. However, action detection requires not only categorizing actions, but also localizing them in untrimmed videos. Therefore, transferring knowledge pertaining to temporal relations is critical for this task which is missing in the previous cross-modal KD frameworks. To this end, we aim at learning an augmented RGB representation for action detection, taking advantage of additional modalities at training time through KD. We propose a KD framework consisting of two levels of distillation. On one hand, atomic-level distillation encourages the RGB student to learn the sub-representation of the actions from the teacher in a contrastive manner. On the other hand, sequence-level distillation encourages the student to learn the temporal knowledge from the teacher, which consists of transferring the Global Contextual Relations and the action Boundary Saliency. The result is an Augmented-RGB stream that can achieve competitive performance as the two-stream network while using only RGB at inference time. Extensive experimental analysis shows that our proposed distillation framework is generic and outperforms other popular cross-modal distillation methods in the action detection task.
****

VaPiD: A Rapid Vanishing Point Detector via Learned Optimizers
Shichen Liu, Yichao Zhou, Yajie Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12859-12868
Being able to infer 3D structures from 2D images with geometric principles, vanishing points have been a well-recognized concept in 3D vision research. It has been widely used in autonomous driving, SLAM, and AR/VR for applications including road direction estimation, camera calibration, and camera pose estimation. Existing vanishing point detection methods often need to trade off between robustness, precision, and inference speed. In this paper, we introduce VaPiD, a novel neural network-based rapid Vanishing Point Detector that achieves unprecedented efficiency with learned vanishing point optimizers. The core of our method contains two components: a vanishing point proposal network that gives a set of vanishing point proposals as coarse estimations; and a neural vanishing point optimizer that iteratively optimizes the positions of the vanishing point proposals to achieve high-precision levels. Extensive experiments on both synthetic and real-world datasets show that our method provides competitive, if not better, performance as compared to the previous state-of-the-art vanishing point detection approaches, while being significantly faster.
****

Deep Survival Analysis With Longitudinal X-Rays for COVID-19
Michelle Shu, Richard Strong Bowen, Charles Herrmann, Gengmo Qi, Michele Santacatterina, Ramin Zabih; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4046-4055
Time-to-event analysis is an important statistical tool for allocating clinical resources such as ICU beds. However, classical techniques like the Cox model cannot directly incorporate images due to their high dimensionality. We propose a deep learning approach that naturally incorporates multiple, time-dependent imaging studies as well as non-imaging data into time-to-event analysis. Our techniques are benchmarked on a clinical dataset of 1,894 COVID-19 patients, and show that image sequences significantly improve predictions. For example, classical time-to-event methods produce a concordance error of around 30-40% for predicting h

ospital admission, while our error is 25% without images and 20% with multiple X-rays included. Ablation studies suggest that our models are not learning spurious features such as scanner artifacts. While our focus and evaluation is on COVID-19, the methods we develop are broadly applicable.
****

Cluster-Promoting Quantization With Bit-Drop for Minimizing Network Quantization Loss
Jung Hyun Lee, Jihun Yun, Sung Ju Hwang, Eunho Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5370-5379
Network quantization, which aims to reduce the bit-lengths of the network weights and activations, has emerged for their deployments to resource-limited devices. Although recent studies have successfully discretized a full-precision network, they still incur large quantization errors after training, thus giving rise to a significant performance gap between a full-precision network and its quantized counterpart. In this work, we propose a novel quantization method for neural networks, Cluster-Promoting Quantization (CPQ) that finds the optimal quantization grids while naturally encouraging the underlying full-precision weights to gather around those quantization grids cohesively during training. This property of CPQ is thanks to our two main ingredients that enable differentiable quantization: i) the use of the categorical distribution designed by a specific probabilistic parametrization in the forward pass and ii) our proposed multi-class straight-through estimator (STE) in the backward pass. Since our second component, multi-class STE, is intrinsically biased, we additionally propose a new bit-drop technique, DropBits, that revises the standard dropout regularization to randomly drop bits instead of neurons. As a natural extension of DropBits, we further introduce the way of learning heterogeneous quantization levels to find proper bit-length for each layer by imposing an additional regularization on DropBits. We experimentally validate our method on various benchmark datasets and network architectures, and also support a new hypothesis for quantization: learning heterogeneous quantization levels outperforms the case using the same but fixed quantization levels from scratch.
****

Continual Prototype Evolution: Learning Online From Non-Stationary Data Streams
Matthias De Lange, Tinne Tuytelaars; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8250-8259
Attaining prototypical features to represent class distributions is well established in representation learning. However, learning prototypes online from streaming data proves a challenging endeavor as they rapidly become outdated, caused by an ever-changing parameter space during the learning process. Additionally, continual learning does not assume the data stream to be stationary, typically resulting in catastrophic forgetting of previous knowledge. As a first, we introduce a system addressing both problems, where prototypes evolve continually in a shared latent space, enabling learning and prediction at any point in time. To facilitate learning, a novel objective function synchronizes the latent space with the continually evolving prototypes. In contrast to the major body of work in continual learning, data streams are processed in an online fashion without task information and can be highly imbalanced, for which we propose an efficient memory scheme. As an additional contribution, we propose the learner-evaluator framework that i) generalizes existing paradigms in continual learning, ii) introduces data incremental learning, and iii) models the bridge between continual learning and concept drift. We obtain state-of-the-art performance by a significant margin on eight benchmarks, including three highly imbalanced data streams. Code is publicly available.
****

Iterative Label Cleaning for Transductive and Semi-Supervised Few-Shot Learning
Michalis Lazarou, Tania Stathaki, Yannis Avrithis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8751-8760
Few-shot learning amounts to learning representations and acquiring knowledge such that novel tasks may be solved with both supervision and data being limited. Improved performance is possible by transductive inference, where the entire tes

t set is available concurrently, and semi-supervised learning, where more unlabe
led data is available. Focusing on these two settings, we introduce a new algori
thm that leverages the manifold structure of the labeled and unlabeled data dist
ribution to predict pseudo-labels, while balancing over classes and using the lo
ss value distribution of a limited-capacity classifier to select the cleanest la
bels, iteratively improving the quality of pseudo-labels. Our solution surpasses
 or matches the state of the art results on four benchmark datasets, namely mini
ImageNet, tieredImageNet, CUB and CIFAR-FS, while being robust over feature spac
e pre-processing and the quantity of available data. The publicly available sour
ce code can be found in https://github.com/MichalisLazarou/iLPC
****

Striking a Balance Between Stability and Plasticity for Class-Incremental Learni
ng
Guile Wu, Shaogang Gong, Pan Li; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2021, pp. 1124-1133
Class-incremental learning (CIL) aims at continuously updating a trained model w
ith new classes (plasticity) without forgetting previously learned old ones (sta
bility). Contemporary studies resort to storing representative exemplars for reh
earsal or preventing consolidated model parameters from drifting, but the former
 requires an additional space for storing exemplars at every incremental phase w
hile the latter usually shows poor model generalization. In this paper, we focus
 on resolving the stability-plasticity dilemma in class-incremental learning whe
re no exemplars from old classes are stored. To make a trade-off between learnin
g new information and maintaining old knowledge, we reformulate a simple yet eff
ective baseline method based on a cosine classifier framework and reciprocal ada
ptive weights. With the reformulated baseline, we present two new approaches to
CIL by learning class-independent knowledge and multi-perspective knowledge, res
pectively. The former exploits class-independent knowledge to bridge learning ne
w and old classes, while the latter learns knowledge from different perspectives
 to facilitate CIL. Extensive experiments on several widely used CIL benchmark d
atasets show the superiority of our approaches over the state-of-the-art methods
.
****

Few-Shot and Continual Learning With Attentive Independent Mechanisms
Eugene Lee, Cheng-Han Huang, Chen-Yi Lee; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2021, pp. 9455-9464
Deep neural networks (DNNs) are known to perform well when deployed to test dist
ributions that shares high similarity with the training distribution. Feeding DN
Ns with new data sequentially that were unseen in the training distribution has
two major challenges --- fast adaptation to new tasks and catastrophic forgettin
g of old tasks. Such difficulties paved way for the on-going research on few-sho
t learning and continual learning. To tackle these problems, we introduce Attent
ive Independent Mechanisms (AIM). We incorporate the idea of learning using fast
 and slow weights in conjunction with the decoupling of the feature extraction a
nd higher-order conceptual learning of a DNN. AIM is designed for higher-order c
onceptual learning, modeled by a mixture of experts that compete to learn indepe
ndent concepts to solve a new task. AIM is a modular component that can be inser
ted into existing deep learning frameworks. We demonstrate its capability for fe
w-shot learning by adding it to SIB and trained on MiniImageNet and CIFAR-FS, sh
owing significant improvement. AIM is also applied to ANML and OML trained on Om
niglot, CIFAR-100 and MiniImageNet to demonstrate its capability in continual le
arning.
****

Trash To Treasure: Harvesting OOD Data With Cross-Modal Matching for Open-Set Se
mi-Supervised Learning
Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei,
Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF International Conference on C
omputer Vision (ICCV), 2021, pp. 8310-8319
Open-set semi-supervised learning (open-set SSL) investigates a challenging but
practical scenario where out-of-distribution (OOD) samples are contained in the

unlabeled data. While the mainstream technique seeks to completely filter out th
e OOD samples for semi-supervised learning (SSL), we propose a novel training me
chanism that could effectively exploit the presence of OOD data for enhanced fea
ture learning while avoiding its adverse impact on the SSL. We achieve this goal
 by first introducing a warm-up training that leverages all the unlabeled data,
including both the in-distribution (ID) and OOD samples. Specifically, we perfor
m a pretext task that enforces our feature extractor to obtain a high-level sema
ntic understanding of the training images, leading to more discriminative featur
es that can benefit the downstream tasks. Since the OOD samples are inevitably d
etrimental to SSL, we propose a novel cross-modal matching strategy to detect OO
D samples. Instead of directly applying binary classification, we train the netw
ork to predict whether the data sample is matched to an assigned one-hot class l
abel. The appeal of the proposed cross-modal matching over binary classification
 is the ability to generate a compatible feature space that aligns with the core
 classification task. Extensive experiments show that our approach substantially
 lifts the performance on open-set SSL and outperforms the state-of-the-art by a
 large margin.
****

AdaFit: Rethinking Learning-Based Normal Estimation on Point Clouds
Runsong Zhu, Yuan Liu, Zhen Dong, Yuan Wang, Tengping Jiang, Wenping Wang, Bishe
ng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision
 (ICCV), 2021, pp. 6118-6127
This paper presents a neural network for robust normal estimation on point cloud
s, named AdaFit, that can deal with point clouds with noise and density variatio
ns. Existing works use a network to learn point-wise weights for weighted least
squares surface fitting to estimate the normals, which has difficulty in finding
 accurate normals in complex regions or containing noisy points. By analyzing th
e step of weighted least squares surface fitting, we find that it is hard to det
ermine the polynomial order of the fitting surface and the fitting surface is se
nsitive to outliers. To address these problems, we propose a simple yet effectiv
e solution that adds an additional offset prediction to improve the quality of n
ormal estimation. Furthermore, in order to take advantage of points from differe
nt neighborhood sizes, a novel Cascaded Scale Aggregation layer is proposed to h
elp the network predict more accurate point-wise offsets and weights. Extensive
experiments demonstrate that AdaFit achieves state-of-the-art performance on bot
h the synthetic PCPNet dataset and the real-word SceneNN dataset.
****

Regularizing Nighttime Weirdness: Efficient Self-Supervised Monocular Depth Esti
mation in the Dark
Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, Jian Yang; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
21, pp. 16055-16064
Monocular depth estimation aims at predicting depth from a single image or video
. Recently, self-supervised methods draw much attention since they are free of d
epth annotations and achieve impressive performance on several daytime benchmark
s. However, they produce weird outputs in more challenging nighttime scenarios b
ecause of low visibility and varying illuminations, which bring weak textures an
d break brightness-consistency assumption, respectively. To address these proble
ms, in this paper we propose a novel framework with several improvements: (1) we
 introduce Priors-Based Regularization to learn distribution knowledge from unpa
ired depth maps and prevent model from being incorrectly trained; (2) we leverag
e Mapping-Consistent Image Enhancement module to enhance image visibility and co
ntrast while maintaining brightness consistency; and (3) we present Statistics-B
ased Mask strategy to tune the number of removed pixels within textureless regio
ns, using dynamic statistics. Experimental results demonstrate the effectiveness
 of each component. Meanwhile, our framework achieves remarkable improvements an
d state-of-the-art results on two nighttime datasets.
****

Occluded Person Re-Identification With Single-Scale Global Representations
Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, Chunhua Shen; Procee

dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11875-11884
Occluded person re-identification (ReID) aims at re-identifying occluded pedestrians from occluded or holistic images taken across multiple cameras. Current state-of-the-art (SOTA) occluded ReID models rely on some auxiliary modules, including pose estimation, feature pyramid and graph matching modules, to learn multi-scale and/or part-level features to tackle the occlusion challenges. This unfortunately leads to complex ReID models that (i) fail to generalize to challenging occlusions of diverse appearance, shape or size, and (ii) become ineffective in handling non-occluded pedestrians. However, real-world ReID applications typically have highly diverse occlusions and involve a hybrid of occluded and non-occluded pedestrians. To address these two issues, we introduce a novel ReID model that learns discriminative single-scale global-level pedestrian features by enforcing a novel exponentially sensitive yet bounded distance loss on occlusion-based augmented data. We show for the first time that learning single-scale global features without using these auxiliary modules is able to outperform those SOTA multi-scale and/or part-level feature-based models. Further, our simple model can achieve new SOTA performance in both occluded and non-occluded ReID, as shown by extensive results on three occluded and two general ReID benchmarks. Additionally, we create a large-scale occluded person ReID dataset with both indoor and outdoor occlusions in different scenes, which is significantly larger and contains substantially more diverse occlusions and pedestrian dressings than existing occluded ReID datasets, providing a more faithful occluded ReID benchmark.
****

Nerfies: Deformable Neural Radiance Fields
Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, Ricardo Martin-Brualla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5865-5874
We present the first method capable of photorealistically reconstructing deformable scenes using photos/videos captured casually from mobile phones. Our approach augments neural radiance fields (NeRF) by optimizing an additional continuous volumetric deformation field that warps each observed point into a canonical 5D NeRF. We observe that these NeRF-like deformation fields are prone to local minima, and propose a coarse-to-fine optimization method for coordinate-based models that allows for more robust optimization. By adapting principles from geometry processing and physical simulation to NeRF-like models, we propose an elastic regularization of the deformation field that further improves robustness. We show that our method can turn casually captured selfie photos/videos into deformable NeRF models that allow for photorealistic renderings of the subject from arbitrary viewpoints, which we dub "nerfies." We evaluate our method by collecting time-synchronized data using a rig with two mobile phones, yielding train/validation images of the same pose at different viewpoints. We show that our method faithfully reconstructs non-rigidly deforming scenes and reproduces unseen views with high fidelity.
****

Towards Novel Target Discovery Through Open-Set Domain Adaptation
Taotao Jing, Hongfu Liu, Zhengming Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9322-9331
Open-set domain adaptation (OSDA) considers that the target domain contains samples from novel categories unobserved in external source domain. Unfortunately, existing OSDA methods always ignore the demand for the information of unseen categories and simply recognize them as "unknown" set without further explanation. This motivates us to understand the unknown categories more specifically by exploring the underlying structures and recovering their interpretable semantic attributes. In this paper, we propose a novel framework to accurately identify the seen categories in target domain, and effectively recover the semantic attributes for unseen categories. Specifically, structure preserving partial alignment is developed to recognize the seen categories through domain-invariant feature learning. Attribute propagation over visual graph is designed to smoothly transit attributes from seen to unseen categories via visual-semantic mapping. Moreover, tw

o new cross-main benchmarks are constructed to evaluate the proposed framework i
n the novel and practical challenge. Experimental results on open-set recognitio
n and semantic recovery demonstrate the superiority of the proposed method over
other compared baselines.
****

Interpretable Visual Reasoning via Induced Symbolic Space
Zhonghao Wang, Kai Wang, Mo Yu, Jinjun Xiong, Wen-mei Hwu, Mark Hasegawa-Johnson
, Humphrey Shi; Proceedings of the IEEE/CVF International Conference on Computer
 Vision (ICCV), 2021, pp. 1878-1887
We study the problem of concept induction in visual reasoning, i.e., identifying
 concepts and their hierarchical relationships from question-answer pairs associ
ated with images; and achieve an interpretable model via working on the induced
symbolic concept space. To this end, we first design a new framework named objec
t-centric compositional attention model (OCCAM) to perform the visual reasoning
task with object-level visual features. Then, we come up with a method to induce
 concepts of objects and relations using clues from the attention patterns betwe
en objects' visual features and question words. Finally, we achieve a higher lev
el of interpretability by imposing OCCAM on the objects represented in the induc
ed symbolic concept space. Experiments on the CLEVR and GQA datasets demonstrate
: 1) our OCCAM achieves a new state of the art without human-annotated functiona
l programs; 2) our induced concepts are both accurate and sufficient as OCCAM ac
hieves an on-par performance on objects represented either in visual features or
 in the induced symbolic concept space.
****

Generalizing Gaze Estimation With Outlier-Guided Collaborative Adaptation
Yunfei Liu, Ruicong Liu, Haofei Wang, Feng Lu; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2021, pp. 3835-3844
Deep neural networks have significantly improved appearance-based gaze estimatio
n accuracy. However, it still suffers from unsatisfactory performance when gener
alizing the trained model to new domains, e.g., unseen environments or persons.
In this paper, we propose a plug-and-play gaze adaptation framework (PnP-GA), wh
ich is an ensemble of networks that learn collaboratively with the guidance of o
utliers. Since our proposed framework does not require ground-truth labels in th
e target domain, the existing gaze estimation networks can be directly plugged i
nto PnP-GA and generalize the algorithms to new domains. We test PnP-GA on four
gaze domain adaptation tasks, ETH-to-MPII, ETH-to-EyeDiap, Gaze360-to-MPII, and
Gaze360-to-EyeDiap. The experimental results demonstrate that the PnP-GA framewo
rk achieves considerable performance improvements of 36.9%, 31.6%, 19.4%, and 11
.8% over the baseline system. The proposed framework also outperforms the state-
of-the-art domain adaptation approaches on gaze domain adaptation tasks.
****

Language-Guided Global Image Editing via Cross-Modal Cyclic Mechanism
Wentao Jiang, Ning Xu, Jiayun Wang, Chen Gao, Jing Shi, Zhe Lin, Si Liu; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 2115-2124
Editing an image automatically via a linguistic request can significantly save l
aborious manual work and is friendly to photography novice. In this paper, we fo
cus on the task of language-guided global image editing. Existing works suffer f
rom imbalanced data distribution of real-world datasets and thus fail to underst
and language requests well. To handle this issue, we propose to create a cycle w
ith our image generator by creating another model called Editing Description Net
work (EDNet) which predicts an editing embedding given a pair of images. Given t
he cycle, we propose several free augmentation strategies to help our model unde
rstand various editing requests given the imbalanced dataset. In addition, two o
ther novel ideas are proposed: an Image-Request Attention (IRA) module which all
ows our method to edit an image spatial-adaptively when the image requires diffe
rent editing degree at different regions, as well as a new evaluation metric for
 this task which is more semantic and reasonable than conventional pixel losses
(eg L1). Extensive experiments on two benchmark datasets demonstrate the effecti
veness of our method over existing approaches.

****
BioFors: A Large Biomedical Image Forensics Dataset
Ekraam Sabir, Soumyaroop Nandi, Wael Abd-Almageed, Prem Natarajan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10963-10973
Research in media forensics has gained traction to combat the spread of misinformation. However, most of this research has been directed towards content generated on social media. Biomedical image forensics is a related problem, where manipulation or misuse of images reported in biomedical research documents is of serious concern. The problem has failed to gain momentum beyond an academic discussion due to an absence of benchmark datasets and standardized tasks. In this paper we present BioFors -- the first dataset for benchmarking common biomedical image manipulations. BioFors comprises 47,805 images extracted from 1,031 open-source research papers. Images in BioFors are divided into four categories -- Microscopy, Blot/Gel, FACS and Macroscopy. We also propose three tasks for forensic analysis -- external duplication detection, internal duplication detection and cut/sharp-transition detection. We benchmark BioFors on all tasks with suitable state-of-the-art algorithms. Our results and analysis show that existing algorithms developed on common computer vision datasets are not robust when applied to biomedical images, validating that more research is required to address the unique challenges of biomedical image forensics.
****
DAM: Discrepancy Alignment Metric for Face Recognition
Jiaheng Liu, Yudong Wu, Yichao Wu, Chuming Li, Xiaolin Hu, Ding Liang, Mengyu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3814-3823
The field of face recognition (FR) has witnessed remarkable progress with the surge of deep learning. The effective loss functions play an important role for FR. In this paper, we observe that a majority of loss functions, including the widespread triplet loss and softmax-based cross-entropy loss, embed inter-class (negative) similarity s_n and intra-class (positive) similarity s_p into similarity pairs and optimize to reduce (s_n - s_p) in the training process. However, in the verification process, existing metrics directly take the absolute similarity between two features as the confidence of belonging to the same identity, which inevitably causes a gap between the training and verification process. To bridge the gap, we propose a new metric called Discrepancy Alignment Metric (DAM) for verification, which introduces the Local Inter-class Discrepancy (LID) for each face image to normalize the absolute similarity score. To estimate the LID of each face image in the verification process, we propose two types of LID Estimation (LIDE) methods, which are reference-based and learning-based estimation methods, respectively. The proposed DAM is plug-and-play and can be easily applied to the most existing methods. Extensive experiments on multiple popular face recognition benchmark datasets demonstrate the effectiveness of our proposed method.
****
Structure-Transformed Texture-Enhanced Network for Person Image Synthesis
Munan Xu, Yuanqi Chen, Shan Liu, Thomas H. Li, Ge Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13859-13868
Pose-guided virtual try-on task aims to modify the fashion item based on pose transfer task. These two tasks that belong to person image synthesis have strong correlations and similarities. However, existing methods treat them as two individual tasks and do not explore correlations between them. Moreover, these two tasks are challenging due to large misalignment and occlusions, thus most of these methods are prone to generate unclear human body structure and blurry fine-grained textures. In this paper, we devise a structure-transformed texture-enhanced network to generate high-quality person images and construct the relationships between two tasks. It consists of two modules: structure-transformed renderer and texture-enhanced stylizer. The structure-transformed renderer is introduced to transform the source person structure to the target one, while the texture-enhanced stylizer is served to enhance detailed textures and controllably inject the fashion style founded on the structural transformation. With the two modules, our

model can generate photorealistic person images in diverse poses and even with various fashion styles. Extensive experiments demonstrate that our approach achieves state-of-the-art results on two tasks.

****

RMSMP: A Novel Deep Neural Network Quantization Framework With Row-Wise Mixed Schemes and Multiple Precisions

Sung-En Chang, Yanyu Li, Mengshu Sun, Weiwen Jiang, Sijia Liu, Yanzhi Wang, Xue Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5251-5260

This work proposes a novel Deep Neural Network (DNN) quantization framework, namely RMSMP, with a \underline R ow-wise \underline M ixed-\underline S cheme and \underline M ulti-\underline P recision approach. Specifically, this is the first effort to assign mixed quantization schemes and multiple precisions within layers -- among rows of the DNN weight matrix, for simplified operations in hardware inference, while preserving accuracy. Furthermore, this paper makes a different observation from the prior work that the quantization error does not necessarily exhibit the layer-wise sensitivity, and actually can be mitigated as long as a certain portion of the weights in every layer are in higher precisions. This observation enables layer-wise uniformality in the hardware implementation towards guaranteed inference acceleration, while still enjoying row-wise flexibility of mixed schemes and multiple precisions to boost accuracy. The candidates of schemes and precisions are derived practically and effectively with a highly hardware-informative strategy to reduce the problem search space. With the offline determined ratio of different quantization schemes and precisions for all the layers, the RMSMP quantization algorithm uses Hessian and variance based method to effectively assign schemes and precisions for each row. The proposed RMSMP is tested for the image classification and natural language processing (BERT) applications, and achieves the best accuracy performance among state-of-the-arts under the same equivalent precisions. The RMSMP is implemented on FPGA devices, achieving 3.65xspeedup in the end-to-end inference time for ResNet-18 on ImageNet, comparing with the 4-bit Fixed-point baseline.

****

Robust Small-Scale Pedestrian Detection With Cued Recall via Memory Learning

Jung Uk Kim, Sungjune Park, Yong Man Ro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3050-3059

Although the visual appearances of small-scale objects are not well observed, humans can recognize them by associating the visual cues of small objects from their memorized appearance. It is called cued recall. In this paper, motivated by the memory process of humans, we introduce a novel pedestrian detection framework that imitates cued recall in detecting small-scale pedestrians. We propose a large-scale embedding learning with the large-scale pedestrian recalling memory (LPR Memory). The purpose of the proposed large-scale embedding learning is to memorize and recall the large-scale pedestrian appearance via the LPR Memory. To this end, we employ the large-scale pedestrian exemplar set, so that, the LPR Memory can recall the information of the large-scale pedestrians from the small-scale pedestrians. Comprehensive quantitative and qualitative experimental results validate the effectiveness of the proposed framework with the LPR Memory.

****

RAIN: Reinforced Hybrid Attention Inference Network for Motion Forecasting

Jiachen Li, Fan Yang, Hengbo Ma, Srikanth Malla, Masayoshi Tomizuka, Chiho Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16096-16106

Motion forecasting plays a significant role in various domains (e.g., autonomous driving, human-robot interaction), which aims to predict future motion sequences given a set of historical observations. However, the observed elements may be of different levels of importance. Some information may be irrelevant or even distracting to the forecasting in certain situations. To address this issue, we propose a generic motion forecasting framework (named RAIN) with dynamic key information selection and ranking based on a hybrid attention mechanism. The general framework is instantiated to handle multi-agent trajectory prediction and human

motion forecasting tasks, respectively. In the former task, the model learns to recognize the relations between agents with a graph representation and to determine their relative significance. In the latter task, the model learns to capture the temporal proximity and dependency in long-term human motions. We also propose an effective double-stage training pipeline with an alternating training strategy to optimize the parameters in different modules of the framework. We validate the framework on both synthetic simulations and motion forecasting benchmarks in different domains, demonstrating that our method not only achieves state-of-the-art forecasting performance but also provides interpretable and reasonable hybrid attention weights.
****

Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images
Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, Faisal Mahmood; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4015-4025
Survival outcome prediction is a challenging weakly-supervised and ordinal regression task in computational pathology that involves modeling complex interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). Despite recent progress in formulating WSIs as bags for multiple instance learning (MIL), representation learning of entire WSIs remains an open and challenging problem, especially in overcoming: 1) the computational complexity of feature aggregation in large bags, and 2) the data heterogeneity gap in incorporating biological priors such as genomic measurements. In this work, we present a Multimodal Co-Attention Transformer (MCAT) framework that learns an interpretable, dense co-attention mapping between WSIs and genomic features formulated in an embedding space. Inspired by approaches in Visual Question Answering (VQA) that can attribute how word embeddings attend to salient objects in an image when answering a question, MCAT learns how histology patches attend to genes when predicting patient survival. In addition to visualizing multimodal interactions, our co-attention transformation also reduces the space complexity of WSI bags, which enables the adaptation of Transformer layers as a general encoder backbone in MIL. We apply our proposed method on five different cancer datasets (4,730 WSIs, 67 million patches). Our experimental results demonstrate that the proposed method consistently achieves superior performance compared to the state-of-the-art methods.
****

AutoShape: Real-Time Shape-Aware Monocular 3D Object Detection
Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, Liangjun Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15641-15650
Existing deep learning-based approaches for monocular 3D object detection in autonomous driving often model the object as a rotated 3D cuboid while the object's geometric shape has been ignored. In this work, we propose an approach for incorporating the shape-aware 2D/3D constraints into the 3D detection framework. Specifically, we employ the deep neural network to learn distinguished 2D keypoints in the 2D image domain and regress their corresponding 3D coordinates in the local 3D object coordinate first. Then the 2D/3D geometric constraints are built by these correspondences for each object to boost the detection performance. For generating the ground truth of 2D/3D keypoints, an automatic model-fitting approach has been proposed by fitting the deformed 3D object model and the object mask in the 2D image. The proposed framework has been verified on the public KITTI dataset and the experimental results demonstrate that by using additional geometrical constraints the detection performance has been significantly improved as compared to the baseline method. More importantly, the proposed framework achieves state-of-the-art performance with real time. Data and code will be available at https://github.com/zongdai/AutoShape
****

Coarsely-Labeled Data for Better Few-Shot Transfer
Cheng Perng Phoo, Bharath Hariharan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9052-9061

Few-shot learning is based on the premise that labels are expensive, especially when they are fine-grained and require expertise. But coarse labels might be eas y to acquire and thus abundant. We present a representation learning approach - PAS that allows few-shot learners to leverage coarsely-labeled data available be fore evaluation. Inspired by self-training, we label the additional data using a teacher trained on the base dataset and filter the teacher's prediction based o n the coarse labels; a new student representation is then trained on the base da taset and the pseudo-labeled dataset. PAS is able to produce a representation th at consistently and significantly outperforms the baselines in 3 different datas ets. Code is available at https://github.com/cpphoo/PAS.
****

Tune It the Right Way: Unsupervised Validation of Domain Adaptation via Soft Nei ghborhood Density
Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, Kate Saenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9184-9193
Unsupervised domain adaptation (UDA) methods can dramatically improve generaliza tion on unlabeled target domains. However, optimal hyper-parameter selection is critical to achieving high accuracy and avoiding negative transfer. Supervised h yper-parameter validation is not possible without labeled target data, which rai ses the question: How can we validate unsupervised adaptation techniques in a re alistic way? We first empirically analyze existing criteria and demonstrate that they are not very effective for tuning hyper-parameters. Intuitively, a well-tr ained source classifier should embed target samples of the same class nearby, fo rming dense neighborhoods in feature space. Based on this assumption, we propose a novel unsupervised validation criterion that measures the density of soft nei ghborhoods by computing the entropy of the similarity distribution between point s. Our criterion is simpler than competing validation methods, yet more effectiv e; it can tune hyper-parameters and the number of training iterations in both im age classification and semantic segmentation models.
****

Improving Neural Network Efficiency via Post-Training Quantization With Adaptive Floating-Point
Fangxin Liu, Wenbo Zhao, Zhezhi He, Yanzhi Wang, Zongwu Wang, Changzhi Dai, Xiao yao Liang, Li Jiang; Proceedings of the IEEE/CVF International Conference on Com puter Vision (ICCV), 2021, pp. 5281-5290
Model quantization has emerged as a mandatory technique for efficient inference with advanced Deep Neural Networks (DNN). It converts the model parameters in fu ll precision (32-bit floating point) to the hardware friendly data representatio n with shorter bit-width, to not only reduce the model size but also simplify th e computation complexity. Nevertheless, prior model quantization either suffers from the inefficient data encoding method thus leading to noncompetitive model c ompression rate, or requires time-consuming quantization aware training process. In this work, we propose a novel Adaptive Floating-Point (AFP) as a variant of standard IEEE-754 floating-point format, with flexible configuration of exponent and mantissa segments. Leveraging the AFP for model quantization (i.e., encodi ng the parameter) could significantly enhance the model compression rate without accuracy degradation and model re-training. We also want to highlight that our proposed AFP could effectively eliminate the computationally intensive de-quant ization step existing in the dynamic quantization technique adopted by the famou s machine learning frameworks (e.g., pytorch, tensorRT and etc). Moreover, we de velop a framework to automatically optimize and choose the adequate AFP configur ation for each layer, thus maximizing the compression efficacy. Our experiments indicate that AFP-encoded ResNet-50/MobileNet-v2 only has ~0.04/0.6% accuracy d egradation w.r.t its full-precision counterpart. It outperforms the state-of-the -art works by 1.1% in accuracy using the same bit-width while reducing the energ y consumption by 11.2x, which is quite impressive for inference.
****

Emerging Properties in Self-Supervised Vision Transformers
Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Boj

anowski, Armand Joulin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9650-9660
In this paper, we question if self-supervised learning provides new properties to Vision Transformer (ViT) that stand out compared to convolutional networks (convnets). Beyond the fact that adapting self-supervised methods to this architecture works particularly well, we make the following observations: first, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent k-NN classifiers, reaching 78.3% top-1 on ImageNet with a small ViT. Our study also underlines the importance of momentum encoder, multi-crop training, and the use of small patches with ViTs. We implement our findings into a simple self-supervised method, called DINO, which we interpret as a form of self-distillation with no labels. We show the synergy between DINO and ViTs by achieving 80.1% top-1 on ImageNet in linear evaluation with ViT-Base.
****

Improving Robustness of Facial Landmark Detection by Defending Against Adversarial Attacks
Congcong Zhu, Xiaoqiang Li, Jide Li, Songmin Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11751-11760
Many recent developments in facial landmark detection have been driven by stacking model parameters or augmenting annotations. However, three subsequent challenges remain, including 1) an increase in computational overhead, 2) the risk of overfitting caused by increasing model parameters, and 3) the burden of labor-intensive annotation by humans. We argue that exploring the weaknesses of the detector so as to remedy them is a promising method of robust facial landmark detection. To achieve this, we propose a sample-adaptive adversarial training (SAAT) approach to interactively optimize an attacker and a detector, which improves facial landmark detection as a defense against sample-adaptive black-box attacks. By leveraging adversarial attacks, the proposed SAAT exploits adversarial perturbations beyond the handcrafted transformations to improve the detector. Specifically, an attacker generates adversarial perturbations to reflect the weakness of the detector. Then, the detector must improve its robustness to adversarial perturbations to defend against adversarial attacks. Moreover, a sample-adaptive weight is designed to balance the risks and benefits of augmenting adversarial examples to train the detector. We also introduce a masked face alignment dataset, Masked-300W, to evaluate our method. Experiments show that our SAAT performed comparably to existing state-of-the-art methods. The dataset and model are publicly available at https://github.com/zhuccly/SAAT.
****

DeepPanoContext: Panoramic 3D Scene Understanding With Holistic Scene Context Graph and Relation-Based Optimization
Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, Yinda Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12632-12641
Panorama images have a much larger field-of-view thus naturally encode enriched scene context information compared to standard perspective images, which however is not well exploited in the previous scene understanding methods. In this paper, we propose a novel method for panoramic 3D scene understanding which recovers the 3D room layout and the shape, pose, position, and semantic category for each object from a single full-view panorama image. In order to fully utilize the rich context information, we design a novel graph neural network based context model to predict the relationship among objects and room layout, and a differentiable relationship-based optimization module to optimize object arrangement with well-designed objective functions on-the-fly. Realizing the existing data are either with incomplete ground truth or overly-simplified scene, we present a new synthetic dataset with good diversity in room layout and furniture placement, and realistic image quality for total panoramic 3D scene understanding. Experiments demonstrate that our method outperforms existing methods on panoramic scene understanding in terms of both geometry accuracy and object arrangement. Code is ava

ilable at https://chengzhag.github.io/publication/dpc.
****

Multi-View Radar Semantic Segmentation
Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, Julien Rebut; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15671-15680
Understanding the scene around the ego-vehicle is key to assisted and autonomous driving. Nowadays, this is mostly conducted using cameras and laser scanners, despite their reduced performances in adverse weather conditions. Automotive radars are low-cost active sensors that measure properties of surrounding objects, including their relative speed, and have the key advantage of not being impacted by rain, snow or fog. However, they are seldom used for scene understanding due to the size and complexity of radar raw data and the lack of annotated datasets. Fortunately, recent open-sourced datasets have opened up research on classification, object detection and semantic segmentation with raw radar signals using end-to-end trainable models. In this work, we propose several novel architectures, and their associated losses, which analyse multiple "views" of the range-angle-Doppler radar tensor to segment it semantically. Experiments conducted on the recent CARRADA dataset demonstrate that our best model outperforms alternative models, derived either from the semantic segmentation of natural images or from radar scene understanding, while requiring significantly fewer parameters. Both our code and trained models are available at https://github.com/valeoai/MVRSS.
****

Exploring Robustness of Unsupervised Domain Adaptation in Semantic Segmentation
Jinyu Yang, Chunyuan Li, Weizhi An, Hehuan Ma, Yuzhi Guo, Yu Rong, Peilin Zhao, Junzhou Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9194-9203
Recent studies imply that deep neural networks are vulnerable to adversarial examples, i.e., inputs with a slight but intentional perturbation are incorrectly classified by the network. Such vulnerability makes it risky for some security-related applications (e.g., semantic segmentation in autonomous cars) and triggers tremendous concerns on the model reliability. For the first time, we comprehensively evaluate the robustness of existing UDA methods and propose a robust UDA approach. It is rooted in two observations: i) the robustness of UDA methods in semantic segmentation remains unexplored, which poses a security concern in this field; and ii) although commonly used self-supervision (e.g., rotation and jigsaw) benefits model robustness in classification and recognition tasks, they fail to provide the critical supervision signals that are essential in semantic segmentation. These observations motivate us to propose adversarial self-supervision UDA (or ASSUDA) that maximizes the agreement between clean images and their adversarial examples by a contrastive loss in the output space. Extensive empirical studies on commonly used benchmarks demonstrate that ASSUDA is resistant to adversarial attacks.
****

Interpretable Image Recognition by Constructing Transparent Embedding Space
Jiaqi Wang, Huafeng Liu, Xinyue Wang, Liping Jing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 895-904
Humans usually explain their reasoning (e.g. classification) by dissecting the image and pointing out the evidence from these parts to the concepts in their minds. Inspired by this cognitive process, several part-level interpretable neural network architectures have been proposed to explain the predictions. However, they suffer from the complex data structure and confusing the effect of the individual part to output category. In this work, an interpretable image recognition deep network is designed by introducing a plug-in transparent embedding space (TesNet) to bridge the high-level input patches (e.g. CNN feature maps) and the output categories. This plug-in embedding space is spanned by transparent basis concepts which are constructed on the Grassmann manifold. These basis concepts are enforced to be category-aware and within-category concepts are orthogonal to each other, which makes sure the embedding space is disentangled. Meanwhile, each basis concept can be traced back to the particular image patches, thus they are t

ransparent and friendly to explain the reasoning process. By comparing with stat e-of-the-art interpretable methods, TesNet is much more beneficial to classifica tion tasks, esp. providing better interpretability on predictions and improve th e final accuracy.

****

Synthesized Feature Based Few-Shot Class-Incremental Learning on a Mixture of Su bspaces

Ali Cheraghian, Shafin Rahman, Sameera Ramasinghe, Pengfei Fang, Christian Simon , Lars Petersson, Mehrtash Harandi; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2021, pp. 8661-8670

Few-shot class incremental learning (FSCIL) aims to incrementally add sets of no vel classes to a well-trained base model in multiple training sessions with the restriction that only a few novel instances are available per class. While learn ing novel classes, FSCIL methods gradually forget base (old) class training and overfit to a few novel class samples. Existing approaches have addressed this pr oblem by computing the class prototypes from the visual or semantic word vector domain. In this paper, we propose addressing this problem using a mixture of sub spaces. Subspaces define the cluster structure of the visual domain and help to describe the visual and semantic domain considering the overall distribution of the data. Additionally, we propose to employ a variational autoencoder (VAE) to generate synthesized visual samples for augmenting pseudo-feature while learning novel classes incrementally. The combined effect of the mixture of subspaces an d synthesized features reduces the forgetting and overfitting problem of FSCIL. Extensive experiments on three image classification datasets show that our propo sed method achieves competitive results compared to state-of-the-art methods.

****

Pyramid Point Cloud Transformer for Large-Scale Place Recognition

Le Hui, Hang Yang, Mingmei Cheng, Jin Xie, Jian Yang; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2021, pp. 6098-6107

Recently, deep learning based point cloud descriptors have achieved impressive r esults in the place recognition task. Nonetheless, due to the sparsity of point clouds, how to extract discriminative local features of point clouds to efficien tly form a global descriptor is still a challenging problem. In this paper, we p ropose a pyramid point cloud transformer network (PPT-Net) to learn the discrimi native global descriptors from point clouds for efficient retrieval. Specificall y, we first develop a pyramid point transformer module that adaptively learns th e spatial relationship of the different local k-NN graphs of point clouds, where the grouped self-attention is proposed to extract discriminative local features of the point clouds. Furthermore, the grouped self-attention not only enhances long-term dependencies of the point clouds, but also reduces the computational c ost. In order to obtain discriminative global descriptors, we construct a pyrami d VLAD module to aggregate the multi-scale feature maps of point clouds into the global descriptors. By applying VLAD pooling on multi-scale feature maps, we ut ilize the context gating mechanism on the multiple global descriptors to adaptiv ely weight the multi-scale global context information into the final global desc riptor. Experimental results on the Oxford dataset and three in-house datasets s how that our method achieves the state-of-the-art on the point cloud based place recognition task. Code is available at https://github.com/fpthink/PPT-Net.

****

Interpreting Attributions and Interactions of Adversarial Attacks

Xin Wang, Shuyun Lin, Hao Zhang, Yufei Zhu, Quanshi Zhang; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1095-1104

This paper aims to explain adversarial attacks in terms of how adversarial pertu rbations contribute to the attacking task. We estimate attributions of different image regions to the decrease of the attacking cost based on the Shapley value. We define and quantify interactions among adversarial perturbation pixels, and decompose the entire perturbation map into relatively independent perturbation c omponents. The decomposition of the perturbation map shows that adversarially-tr ained DNNs have more perturbation components in the foreground than normally-tra ined DNNs. Moreover, compared to the normally-trained DNN, the adversarially-tra

ined DNN have more components which mainly decrease the score of the true categ
ry. Above analyses provide new insights into the understanding of adversarial at
tacks.
****

Neural Photofit: Gaze-Based Mental Image Reconstruction
Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, Andreas Bulli
ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2021, pp. 245-254
We propose a novel method that leverages human fixations to visually decode the
image a person has in mind into a photofit (facial composite). Our method combin
es three neural networks: An encoder, a scoring network, and a decoder. The enco
der extracts image features and predicts a neural activation map for each face l
ooked at by a human observer. A neural scoring network compares the human and ne
ural attention and predicts a relevance score for each extracted image feature.
Finally, image features are aggregated into a single feature vector as a linear
combination of all features weighted by relevance which a decoder decodes into t
he final photofit. We train the neural scoring network on a novel dataset contai
ning gaze data of 19 participants looking at collages of synthetic faces. We sho
w that our method significantly outperforms a mean baseline predictor and report
 on a human study that shows that we can decode photofits that are visually plau
sible and close to the observer's mental image.
****

Efficient and Differentiable Shadow Computation for Inverse Problems
Linjie Lyu, Marc Habermann, Lingjie Liu, Mallikarjun B R, Ayush Tewari, Christia
n Theobalt; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2021, pp. 13107-13116
Differentiable rendering has received increasing interest in the solution of ima
ge-based inverse problems. It can benefit traditional optimization-based solutio
ns to inverse problems, but also allows for self-supervision of learning-based a
pproaches for which training data with ground truth annotation is hard to obtain
. However, existing differentiable renderers either do not correctly model compl
ex visibility responsible for shadows in the images, or are too slow for being u
sed to train deep architectures over thousands of iterations. To this end, we pr
opose an accurate yet efficient approach for differentiable visibility and soft
shadow computation. Our approach is based on the spherical harmonics approximati
on of the scene illumination and visibility, where the occluding surface is appr
oximated with spheres. This allows for significantly more efficient visibility c
omputation compared to methods based on path tracing without sacrificing quality
 of generated images. As our formulation is differentiable, it can be used to so
lve various image-based inverse problems such as texture, lighting, geometry rec
overy from images using analysis-by-synthesis optimization.
****

Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in
 Training Data
Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, Mario Fritz; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14448-1445
7
Photorealistic image generation has reached a new level of quality due to the br
eakthroughs of generative adversarial networks (GANs). Yet, the dark side of suc
h deepfakes, the malicious use of generated media, raises concerns about visual
misinformation. While existing research work on deepfake detection demonstrates
high accuracy, it is subject to advances in generation techniques and adversaria
l iterations on detection countermeasure techniques. Thus, we seek a proactive a
nd sustainable solution on deepfake detection, that is agnostic to the evolution
 of generative models, by introducing artificial fingerprints into the models. O
ur approach is simple and effective. We first embed artificial fingerprints into
 training data, then validate a surprising discovery on the transferability of s
uch fingerprints from training data to generative models, which in turn appears
in the generated deepfakes. Experiments show that our fingerprinting solution (1
) holds for a variety of cutting-edge generative models, (2) leads to a negligib

le side effect on generation quality, (3) stays robust against image-level and m odel-level perturbations, (4) stays hard to be detected by adversaries, and (5) converts deepfake detection and attribution into trivial tasks and outperforms t he recent state-of-the-art baselines. Our solution closes the responsibility loo p between publishing pre-trained generative model inventions and their possible misuses, which makes it independent of the current arms race.
****

TokenPose: Learning Keypoint Tokens for Human Pose Estimation
Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, Erj in Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11313-11322
Human pose estimation deeply relies on visual clues and anatomical constraints b etween parts to locate keypoints. Most existing CNN-based methods do well in vis ual representation, however, lacking in the ability to explicitly learn the cons traint relationships between keypoints. In this paper, we propose a novel approa ch based on Token representation for human Pose estimation (TokenPose). In detai l, each keypoint is explicitly embedded as a token to simultaneously learn const raint relationships and appearance cues from images. Extensive experiments show that the small and large TokenPose models are on par with state-of-the-art CNN-b ased counterparts while being more lightweight. Specifically, our TokenPose-S an d TokenPose-L achieve 72.5 AP and 75.8 AP on COCO validation dataset respectivel y, with significant reduction in parameters and GFLOPs. Code is publicly availab le at https://github.com/leeyegy/TokenPose.
****

Disentangled Lifespan Face Synthesis
Sen He, Wentong Liao, Michael Ying Yang, Yi-Zhe Song, Bodo Rosenhahn, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3877-3886
A lifespan face synthesis (LFS) model aims to generate a set of photo-realistic face images of a person's whole life, given only one snapshot as reference. The generated face image given a target age code is expected to be age-sensitive ref lected by bio-plausible transformations of shape and texture, while being identi ty preserving. This is extremely challenging because the shape and texture chara cteristics of a face undergo separate and highly nonlinear transformations w.r.t . age. Most recent LFS models are based on generative adversarial networks (GANs) whereby age code conditional transformations are applied to a latent face repr esentation. They benefit greatly from the recent advancements of GANs. However, without explicitly disentangling their latent representations into the texture, shape and identity factors, they are fundamentally limited in modeling the nonli near age-related transformation on texture and shape whilst preserving identity. In this work, a novel LFS model is proposed to disentangle the key face charact eristics including shape, texture and identity so that the unique shape and text ure age transformations can be modeled effectively. This is achieved by extracti ng shape, texture and identity features separately from an encoder. Critically, two transformation modules, one conditional convolution based and the other chan nel attention based, are designed for modeling the nonlinear shape and texture f eature transformations respectively. This is to accommodate their rather distinc t aging processes and ensure that our synthesized images are both age-sensitive and identity preserving. Extensive experiments show that our LFS model is clearl y superior to the state-of-the-art alternatives.
****

Dual Transfer Learning for Event-Based End-Task Prediction via Pluggable Event t o Image Translation
Lin Wang, Yujeong Chae, Kuk-Jin Yoon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2135-2145
Event cameras are novel sensors that perceive the per-pixel intensity changes an d output asynchronous event streams with high dynamic range and less motion blur . It has been shown that events alone can be used for end-task learning, e.g., s emantic segmentation, based on encoder-decoder-like networks. However, as events are sparse and mostly reflect edge information, it is difficult to recover orig

inal details merely relying on the decoder. Moreover, most methods resort to the pixel-wise loss alone for supervision, which might be insufficient to fully exploit the visual details from sparse events, thus leading to less optimal performance. In this paper, we propose a simple yet flexible two-stream framework named Dual Transfer Learning (DTL) to effectively enhance the performance on the end-tasks without adding extra inference cost. The proposed approach consists of three parts: event to end-task learning (EEL) branch, event to image translation (EIT) branch, and transfer learning (TL) module that simultaneously explores the feature-level affinity information and pixel-level knowledge from the EIT branch to improve the EEL branch. This simple yet novel method leads to strong representation learning from events and is evidenced by the significant performance boost on the end-tasks such as semantic segmentation and depth estimation.

****

Exploration and Estimation for Model Compression

Yanfu Zhang, Shangqian Gao, Heng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 487-496

Deep neural networks achieve great success in many visual recognition tasks. However, the model deployment is usually subject to some computational resources. Model pruning under computational budget has attracted growing attention. In this paper, we focus on the discrimination-aware compression of Convolutional Neural Networks (CNNs). In prior arts, directly searching the optimal sub-network is an integer programming problem, which is non-smooth, non-convex, and NP-hard. Meanwhile, the heuristic pruning criterion lacks clear interpretability and doesn't generalize well in applications. To address this problem, we formulate sub-networks as samples from a multivariate Bernoulli distribution and resort to the approximation of continuous problem. We propose a new flexible search scheme via alternating exploration and estimation. In the exploration step, we employ stochastic gradient Hamiltonian Monte Carlo with budget-awareness to generate sub-networks, which allows large search space with efficient computation. In the estimation step, we deduce the sub-network sampler to a near-optimal point, to promote the generation of high-quality sub-networks. Unifying the exploration and estimation, our approach avoids early falling into local minimum via a fast gradient-based search in a larger space. Extensive experiments on CIFAR-10 and ImageNet show that our method achieves state-of-the-art performances on pruning several popular CNNs.

****

Task Switching Network for Multi-Task Learning

Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popovi■, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8291-8300

We introduce Task Switching Networks (TSNs), a task-conditioned architecture with a single unified encoder/decoder for efficient multi-task learning. Multiple tasks are performed by switching between them, performing one task at a time. TSNs have a constant number of parameters irrespective of the number of tasks. This scalable yet conceptually simple approach circumvents the overhead and intricacy of task-specific network components in existing works. In fact, we demonstrate for the first time that multi-tasking can be performed with a single task-conditioned decoder. We achieve this by learning task-specific conditioning parameters through a jointly trained task embedding network, encouraging constructive interaction between tasks. Experiments validate the effectiveness of our approach, achieving state-of-the-art results on two challenging multi-task benchmarks, PASCAL-Context and NYUD. Our analysis of the learned task embeddings further indicates a connection to task relationships studied in the recent literature.

****

Interaction Compass: Multi-Label Zero-Shot Learning of Human-Object Interactions via Spatial Relations

Dat Huynh, Ehsan Elhamifar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8472-8483

We study the problem of multi-label zero-shot recognition in which labels are in the form of human-object interactions (combinations of actions on objects), eac

h image may contain multiple interactions and some interactions do not have trai ning images. We propose a novel compositional learning framework that decouples interaction labels into separate action and object scores that incorporate the s patial compatibility between the two components. We combine these scores to effi ciently recognize seen and unseen interactions. However, learning action-object spatial relations, in principle, requires bounding-box annotations, which are co stly to gather. Moreover, it is not clear how to generalize spatial relations to unseen interactions. We address these challenges by developing a cross-attentio n mechanism that localizes objects from action locations and vice versa by predi cting displacements between them, referred to as relational directions. During t raining, we estimate the relational directions as ones maximizing the scores of ground-truth interactions that guide predictions toward compatible action-object regions. By extensive experiments, we show the effectiveness of our framework, where we improve the state of the art by 2.6% mAP score and 5.8% recall score on HICO and Visual Genome datasets, respectively. Code is available at https://git hub.com/hbdat/iccv21_relational_direction.
****

Unsupervised Point Cloud Pre-Training via Occlusion Completion
Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, Matt J. Kusner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9782-9792
We describe a simple pre-training approach for point clouds. It works in three s teps: 1. Mask all points occluded in a camera view; 2. Learn an encoder-decoder model to reconstruct the occluded points; 3. Use the encoder weights as initiali sation for downstream point cloud tasks. We find that even when we pre-train on a single dataset (ModelNet40), this method improves accuracy across different da tasets and encoders, on a wide range of downstream tasks. Specifically, we show that our method outperforms previous pre-training methods in object classificati on, and both part-based and semantic segmentation tasks. We study the pre-traine d features and find that they lead to wide downstream minima, have high transfor mation invariance, and have activations that are highly correlated with part lab els. Code and data are available at https://github.com/hansen7/OcCo
****

Structure-From-Sherds: Incremental 3D Reassembly of Axially Symmetric Pots From Unordered and Mixed Fragment Collections
Je Hyeong Hong, Seong Jong Yoo, Muhammad Arshad Zeeshan, Young Min Kim, Jinwook Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2021, pp. 5443-5451
Re-assembling multiple pots accurately from numerous 3D scanned fragments remain s a challenging task to this date. Previous methods extract all potential matchi ng pairs of pot sherds and considers them simultaneously to search for an optima l global pot configuration. In this work, we empirically show such global approa ch greatly suffers from false positive matches between sherds inflicted by indis tinctive sharp fracture surfaces in pot fragments. To mitigate this problem, we take inspirations from the field of structure-from-motion (SfM), where many pipe lines have matured in reconstructing a 3D scene from multiple images. Motivated by the success of the incremental approach in robust SfM, we present an efficien t reassembly method for axially symmetric pots based on iterative registration o f one sherd at a time. Our method goes beyond replicating incremental SfM and ad dresses indistinguishable false matches by embracing beam search to explore mult itudes of registration possibilities. Additionally, we utilize multiple roots in each step to allow simultaneous reassembly of multiple pots. The proposed appro ach shows above 80% reassembly accuracy on a dataset of real 80 fragments mixed from 5 pots, pushing the state-of-the-art and paving the way towards the goal of large-scale pot reassembly. Our code and preprocessed data will be made availab le for research.
****

Towards Vivid and Diverse Image Colorization With Generative Color Prior
Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, Ying Shan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1437

7-14386
Colorization has attracted increasing interest in recent years. Classic referenc e-based methods usually rely on external color images for plausible results. A l arge image database or online search engine is inevitably required for retrievin g such exemplars. Recent deep-learning-based methods could automatically coloriz e images at a low cost. However, unsatisfactory artifacts and incoherent colors are always accompanied. In this work, we aim at recovering vivid colors by lever aging the rich and diverse color priors encapsulated in a pretrained Generative Adversarial Networks (GAN). Specifically, we first "retrieve" matched features (similar to exemplars) via a GAN encoder and then incorporate these features into the colorization process with feature modulations. Thanks to the powerful gener ative color prior and delicate designs, our method could produce vivid colors wi th a single forward pass. Moreover, it is highly convenient to obtain diverse re sults by modifying GAN latent codes. Our method also inherits the merit of inter pretable controls of GANs and could attain controllable and smooth transitions b y walking through GAN latent space. Extensive experiments and user studies demon strate that our method achieves superior performance than previous works.
****

Asymmetric Loss for Multi-Label Classification
Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Pr otter, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2021, pp. 82-91
In a typical multi-label setting, a picture contains on average few positive lab els, and many negative ones. This positive-negative imbalance dominates the opti mization process, and can lead to under-emphasizing gradients from positive labe ls during training, resulting in poor accuracy. In this paper, we introduce a no vel asymmetric loss (""ASL""), which operates differently on positive and negati ve samples. The loss enables to dynamically down-weights and hard-thresholds eas y negative samples, while also discarding possibly mislabeled samples. We demons trate how ASL can balance the probabilities of different samples, and how this b alancing is translated to better mAP scores. With ASL, we reach state-of-the-art results on multiple popular multi-label datasets: MS-COCO, Pascal-VOC, NUS-WIDE and Open Images. We also demonstrate ASL applicability for other tasks, such as single-label classification and object detection. ASL is effective, easy to imp lement, and does not increase the training time or complexity. Implementation is available at: https://github.com/Alibaba-MIIL/ASL.
****

The Pursuit of Knowledge: Discovering and Localizing Novel Categories Using Dual Memory
Sai Saketh Rambhatla, Rama Chellappa, Abhinav Shrivastava; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9153-9163
We tackle object category discovery, which is the problem of discovering and loc alizing novel objects in a large unlabeled dataset. While existing methods show results on datasets with less cluttered scenes and fewer object instances per im age, we present our results on the challenging COCO dataset. Moreover, we argue that, rather than discovering new categories from scratch, discovery algorithms can benefit from identifying what is already known and focusing their attention on the unknown. We propose a method that exploits prior knowledge about certain object types to discover new categories by leveraging two memory modules, namely Working and Semantic memory. We show the performance of our detector on the COC O minival dataset to demonstrate its in-the-wild capabilities.
****

Unconditional Scene Graph Generation
Sarthak Garg, Helisa Dhamo, Azade Farshad, Sabrina Musatian, Nassir Navab, Feder ico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vi sion (ICCV), 2021, pp. 16362-16371
Despite recent advancements in single-domain or single-object image generation, it is still challenging to generate complex scenes containing diverse, multiple objects and their interactions. Scene graphs, composed of nodes as objects and d irected-edges as relationships among objects, offer an alternative representatio

n of a scene that is more semantically grounded than images. We hypothesize that a generative model for scene graphs might be able to learn the underlying semantic structure of real-world scenes more effectively than images, and hence, generate realistic novel scenes in the form of scene graphs. In this work, we explore a new task for the unconditional generation of semantic scene graphs. We develop a deep auto-regressive model called SceneGraphGen which can directly learn the probability distribution over labelled and directed graphs using a hierarchical recurrent architecture. The model takes a seed object as input and generates a scene graph in a sequence of steps, each step generating an object node, followed by a sequence of relationship edges connecting to the previous nodes. We show that the scene graphs generated by SceneGraphGen are diverse and follow the semantic patterns of real-world scenes. Additionally, we demonstrate the application of the generated graphs in image synthesis, anomaly detection and scene graph completion.

****

Unified Graph Structured Models for Video Understanding

Anurag Arnab, Chen Sun, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8117-8126

Accurate video understanding involves reasoning about the relationships between actors, objects and their environment, often over long temporal intervals. In this paper, we propose a message passing graph neural network that explicitly models these spatio-temporal relations and can use explicit representations of objects, when supervision is available, and implicit representations otherwise. Our formulation generalises previous structured models for video understanding, and allows us to study how different design choices in graph structure and representation affect the model's performance. We demonstrate our method on two different tasks requiring relational reasoning in videos -- spatio-temporal action detection on AVA and UCF101-24, and video scene graph classification on the recent Action Genome dataset -- and achieve state-of-the-art results on all three datasets. Furthermore, we show quantitatively and qualitatively how our method is able to more effectively model relationships between relevant entities in the scene.

****

Minimal Cases for Computing the Generalized Relative Pose Using Affine Correspondences

Banglei Guan, Ji Zhao, Daniel Barath, Friedrich Fraundorfer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6068-6077

We propose three novel solvers for estimating the relative pose of a multi-camera system from affine correspondences (ACs). A new constraint is derived interpreting the relationship of ACs and the generalized camera model. Using the constraint, we demonstrate efficient solvers for two types of motions assumed. Considering that the cameras undergo planar motion, we propose a minimal solution using a single AC and a solver with two ACs to overcome the degenerate case. Also, we propose a minimal solution using two ACs with known vertical direction, e.g., from an IMU. Since the proposed methods require significantly fewer correspondences than state-of-the-art algorithms, they can be efficiently used within RANSAC for outlier removal and initial motion estimation. The solvers are tested both on synthetic data and on real-world scenes from the KITTI odometry benchmark. It is shown that the accuracy of the estimated poses is superior to the state-of-the-art techniques.

****

Towards Efficient Graph Convolutional Networks for Point Cloud Handling

Yawei Li, He Chen, Zhaopeng Cui, Radu Timofte, Marc Pollefeys, Gregory S. Chirikjian, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3752-3762

We aim at improving the computational efficiency of graph convolutional networks (GCNs) for learning on point clouds. The basic graph convolution that is composed of a K-nearest neighbor (KNN) search and a multilayer perceptron (MLP) is examined. By mathematically analyzing the operations there, two findings to improve the efficiency of GCNs are obtained. (1) The local geometric structure information of 3D representations propagates smoothly across the GCN that relies on KNN

search to gather neighborhood features. This motivates the simplification of multiple KNN searches in GCNs. (2) Shuffling the order of graph feature gathering and an MLP leads to equivalent or similar composite operations. Based on those findings, we optimize the computational procedure in GCNs. A series of experiments show that the optimized networks have reduced computational complexity, decreased memory consumption, and accelerated inference speed while maintaining comparable accuracy for learning on point clouds.

****

Gait Recognition in the Wild: A Benchmark

Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14789-14799

Gait benchmarks empower the research community to train and evaluate high-performance gait recognition systems. Even though growing efforts have been devoted to cross-view recognition, academia is restricted by current existing databases captured in the controlled environment. In this paper, we contribute a new benchmark for Gait REcognition in the Wild (GREW). The GREW dataset is constructed from natural videos, which contains hundreds of cameras and thousands of hours streams in open systems. With tremendous manual annotations, the GREW consists of 26K identities and 128K sequences with rich attributes for unconstrained gait recognition. Moreover, we add a distractor set of over 233K sequences, making it more suitable for real-world applications. Compared with prevailing predefined cross-view datasets, the GREW has diverse and practical view variations, as well as more natural challenging factors. To the best of our knowledge, this is the first large-scale dataset for gait recognition in the wild. Equipped with this benchmark, we dissect the unconstrained gait recognition problem. Representative appearance-based and model-based methods are explored, and comprehensive baselines are established. Experimental results show (1) The proposed GREW benchmark is necessary for training and evaluating gait recognizer in the wild. (2) For state-of-the-art gait recognition approaches, there is a lot of room for improvement. (3) The GREW benchmark can be used as effective pre-training for controlled gait recognition. Benchmark website is https://www.grew-benchmark.org/.

****

Structured Bird's-Eye-View Traffic Scene Understanding From Onboard Images

Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15661-15670

Autonomous navigation requires structured representation of the road network and instance-wise identification of the other traffic agents. Since the traffic scene is defined on the ground plane, this corresponds to scene understanding in the bird's-eye-view (BEV). However, the onboard cameras of autonomous cars are customarily mounted horizontally for a better view of the surrounding, making this task very challenging. In this work, we study the problem of extracting a directed graph representing the local road network in BEV coordinates, from a single onboard camera image. Moreover, we show that the method can be extended to detect dynamic objects on the BEV plane. The semantics, locations, and orientations of the detected objects together with the road graph facilitates a comprehensive understanding of the scene. Such understanding becomes fundamental for the downstream tasks, such as path planning and navigation. We validate our approach against powerful baselines and show that our network achieves superior performance. We also demonstrate the effects of various design choices through ablation studies.

****

MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?

Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, Rita Cucchiara; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10849-10859

Deep learning-based methods for video pedestrian detection and tracking require large volumes of training data to achieve good performance. However, data acquis

ition in crowded public environments raises data privacy concerns -- we are not allowed to simply record and store data without the explicit consent of all participants. Furthermore, the annotation of such data for computer vision applications usually requires a substantial amount of manual effort, especially in the video domain. Labeling instances of pedestrians in highly crowded scenarios can be challenging even for human annotators and may introduce errors in the training data. In this paper, we study how we can advance different aspects of multi-person tracking using solely synthetic data. To this end, we generate MOTSynth, a large, highly diverse synthetic dataset for object detection and tracking using a rendering game engine. Our experiments show that MOTSynth can be used as a replacement for real data on tasks such as pedestrian detection, re-identification, segmentation, and tracking.

****

MonteFloor: Extending MCTS for Reconstructing Accurate Large-Scale Floor Plans
Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, Vincent Lepetit; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16034-16043

We propose a novel method for reconstructing floor plans from noisy 3D point clouds. Our main contribution is a principled approach that relies on the Monte Carlo Tree Search (MCTS) algorithm to maximize a suitable objective function efficiently despite the complexity of the problem. Like previous work, we first project the input point cloud to a top view to create a density map and extract room proposals from it. Our method selects and optimizes the polygonal shapes of these room proposals jointly to fit the density map and outputs an accurate vectorized floor map even for large complex scenes. To do this, we adapted MCTS, an algorithm originally designed to learn to play games, to select the room proposals by maximizing an objective function combining the fitness with the density map as predicted by a deep network and regularizing terms on the room shapes. We also introduce a refinement step to MCTS that adjusts the shape of the room proposals. For this step, we propose a novel differentiable method for rendering the polygonal shapes of these proposals. We evaluate our method on the recent and challenging Structured3D and Floor-SP datasets and show a significant improvement over the state-of-the-art, without imposing any hard constraints nor assumptions on the floor plan configurations.

****

Relaxed Transformer Decoders for Direct Action Proposal Generation
Jing Tan, Jiaqi Tang, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13526-13535

Temporal action proposal generation is an important and challenging task in video understanding, which aims at detecting all temporal segments containing action instances of interest. The existing proposal generation approaches are generally based on pre-defined anchor windows or heuristic bottom-up boundary matching strategies. This paper presents a simple and efficient framework (RTD-Net) for direct action proposal generation, by re-purposing a Transformer-alike architecture. To tackle the essential visual difference between time and space, we make three important improvements over the original transformer detection framework (DETR). First, to deal with slowness prior in videos, we replace the original Transformer encoder with a boundary attentive module to better capture long-range temporal information. Second, due to the ambiguous temporal boundary and relatively sparse annotations, we present a relaxed matching scheme to relieve the strict criteria of single assignment to each groundtruth. Finally, we devise a three-branch head to further improve the proposal confidence estimation by explicitly predicting its completeness. Extensive experiments on THUMOS14 and ActivityNet-1.3 benchmarks demonstrate the effectiveness of RTD-Net, on both tasks of temporal action proposal generation and temporal action detection. Moreover, due to its simplicity in design, our framework is more efficient than previous proposal generation methods, without non-maximum suppression post-processing. The code and models are made available at https://github.com/MCG-NJU/RTD-Action.

****

D2-Net: Weakly-Supervised Action Localization via Discriminative Embeddings and

Denoised Activations
Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13608-13617
This work proposes a weakly-supervised temporal action localization framework, called D2-Net, which strives to temporally localize actions using video-level supervision. Our main contribution is the introduction of a novel loss formulation, which jointly enhances the discriminability of latent embeddings and robustness of the output temporal class activations with respect to foreground-background noise caused by weak supervision. The proposed formulation comprises a discriminative and a denoising loss term for enhancing temporal action localization. The discriminative term incorporates a classification loss and utilizes a top-down attention mechanism to enhance the separability of latent foreground-background embeddings. The denoising loss term explicitly addresses the foreground-background noise in class activations by simultaneously maximizing intra-video and inter-video mutual information using a bottom-up attention mechanism. As a result, activations in the foreground regions are emphasized whereas those in the background regions are suppressed, thereby leading to more robust predictions. Comprehensive experiments are performed on multiple benchmarks, including THUMOS14 and ActivityNet1.2. Our D2-Net performs favorably in comparison to the existing methods on all datasets, achieving gains as high as 2.3% in terms of mAP at IoU=0.5 on THUMOS14. Source code is available at https://github.com/naraysa/D2-Net.
****
Auto Graph Encoder-Decoder for Neural Network Pruning
Sixing Yu, Arya Mazaheri, Ali Jannesari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6362-6372
Model compression aims to deploy deep neural networks (DNN) on mobile devices with limited computing and storage resources. However, most of the existing model compression methods rely on manually defined rules, which require domain expertise. DNNs are essentially computational graphs, which contain rich structural information. In this paper, we aim to find a suitable compression policy from DNNs' structural information. We propose an automatic graph encoder-decoder model compression (AGMC) method combined with graph neural networks (GNN) and reinforcement learning (RL). We model the target DNN as a graph and use GNN to learn the DNN's embeddings automatically. We compared our method with rule-based DNN embedding model compression methods to show the effectiveness of our method. Results show that our learning-based DNN embedding achieves better performance and a higher compression ratio with fewer search steps. We evaluated our method on over-parameterized and mobile-friendly DNNs and compared our method with handcrafted and learning-based model compression approaches. On over parameterized DNNs, such as ResNet-56, our method outperformed handcrafted and learning-based methods with 4.36% and 2.56% higher accuracy, respectively. Furthermore, on MobileNet-v2, we achieved a higher compression ratio than state-of-the-art methods with just 0.93% accuracy loss.
****
Adaptive Surface Reconstruction With Multiscale Convolutional Kernels
Benjamin Ummenhofer, Vladlen Koltun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5651-5660
We propose generalized convolutional kernels for 3D reconstruction with ConvNets from point clouds. Our method uses multiscale convolutional kernels that can be applied to adaptive grids as generated with octrees. In addition to standard kernels in which each element has a distinct spatial location relative to the center, our elements have a distinct relative location as well as a relative scale level. Making our kernels span multiple resolutions allows us to apply ConvNets to adaptive grids for large problem sizes where the input data is sparse but the entire domain needs to be processed. Our ConvNet architecture can predict the signed and unsigned distance fields for large data sets with millions of input points and is faster and more accurate than classic energy minimization or recent learning approaches. We demonstrate this in a zero-shot setting where we only train on synthetic data and evaluate on the Tanks and Temples dataset of real-world

large-scale 3D scenes.
****
Localized Simple Multiple Kernel K-Means
Xinwang Liu, Sihang Zhou, Li Liu, Chang Tang, Siwei Wang, Jiyuan Liu, Yi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9293-9301
As a representative of multiple kernel clustering (MKC), simple multiple kernel k-means (SimpleMKKM) is recently put forward to boosting the clustering performance by optimally fusing a group of pre-specified kernel matrices. Despite achieving significant improvement in a variety of applications, we find out that SimpleMKKM could indiscriminately force all sample pairs to be equally aligned with the same ideal similarity. As a result, it does not sufficiently take the variation of samples into consideration, leading to unsatisfying clustering performance. To address these issues, this paper proposes a novel MKC algorithm with a "local" kernel alignment, which only requires that the similarity of a sample to its k-nearest neighbours be aligned with the ideal similarity matrix. Such an alignment helps the clustering algorithm to focus on closer sample pairs that shall stay together and avoids involving unreliable similarity evaluation for farther sample pairs. After that, we theoretically show that the objective of SimpleMKKM is a special case of this local kernel alignment criterion with normalizing each base kernel matrix. Based on this observation, the proposed localized SimpleMKKM can be readily implemented by existing SimpleMKKM package. Moreover, we conduct extensive experiments on several widely used benchmark datasets to evaluate the clustering performance of localized SimpleMKKM. The experimental results have demonstrated that our algorithm consistently outperforms the state-of-the-art ones, verifying the effectiveness of the proposed local kernel alignment criterion. The code of Localized SimpleMKKM is publicly available at: https://github.com/xinwangliu/LocalizedSMKKM.
****
SmartShadow: Artistic Shadow Drawing Tool for Line Drawings
Lvmin Zhang, Jinyue Jiang, Yi Ji, Chunping Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5391-5400
SmartShadow is a deep learning application for digital painting artists to draw shadows on line drawings, with three proposed tools. (1) Shadow brush: artists can draw scribbles to coarsely indicate the areas inside or outside their wanted shadows, and the application will generate the shadows in real-time. (2) Shadow boundary brush: this brush can precisely control the boundary of any specific shadow. (3) Global shadow generator: this tool can estimate the global shadow direction from input brush scribbles, and then consistently propagate local shadows to the entire image. These three tools can not only speed up the shadow drawing process (by 3.1 times as experiments validate), but also allow for the flexibility to achieve various shadow effects and facilitate richer artistic creations. To this end, we train Convolutional Neural Networks (CNNs) with a collected large-scale dataset of both real and synthesized data, and especially, we collect 1670 shadow samples drawn by real artists. Both qualitative analysis and user study show that our approach can generate high-quality shadows that are practically usable in the daily works of digital painting artists. We present 30 additional results and 15 visual comparisons in the supplementary materiel.
****
PT-CapsNet: A Novel Prediction-Tuning Capsule Network Suitable for Deeper Architectures
Chenbin Pan, Senem Velipasalar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11996-12005
Capsule Networks (CapsNets) create internal representations by parsing inputs into various instances at different resolution levels via a two-phase process -- part-whole transformation and hierarchical component routing. Since both of these internal phases are computationally expensive, CapsNets have not found wider use. Existing variations of CapsNets mainly focus on performance comparison with the original CapsNet, and have not outperformed CNN-based models on complex tasks. To address the limitations of the existing CapsNet structures, we propose a no

vel Prediction-Tuning Capsule Network (PT-CapsNet), and also introduce fully connected PT-Capsules (FC-PT-Caps) and locally connected PT-Capsules (LC-PT-Caps). Different from existing CapsNet structures, our proposed model (i) allows the use of capsules for more difficult vision tasks and provides wider applicability; and (ii) provides better than or comparable performance to CNN-based baselines on these complex tasks. In our experiments, we show robustness to affine transformations, as well as the lightweight and scalability of PT-CapsNet via constructing larger and deeper networks and performing comparisons on classification, semantic segmentation and object detection tasks. The results show consistent performance improvement and significant parameter reduction compared to various baseline models. Code is available at https://github.com/Christinepan881/PT-CapsNet.git.

****

In-Place Scene Labelling and Understanding With Implicit Scene Representation
Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15838-15847

Semantic labelling is highly correlated with geometry and radiance reconstruction, as scene entities with similar shape and appearance are more likely to come from similar classes. Recent implicit neural reconstruction techniques are appealing as they do not require prior training data, but the same fully self-supervised approach is not possible for semantics because labels are human-defined properties. We extend neural radiance fields (NeRF) to jointly encode semantics with appearance and geometry, so that complete and accurate 2D semantic labels can be achieved using a small amount of in-place annotations specific to the scene. The intrinsic multi-view consistency and smoothness of NeRF benefit semantics by enabling sparse labels to efficiently propagate. We show the benefit of this approach when labels are either sparse or very noisy in room-scale scenes. We demonstrate its advantageous properties in various interesting applications such as an efficient scene labelling tool, novel semantic view synthesis, label denoising, super-resolution, label interpolation and multi-view semantic label fusion in visual semantic mapping systems.

****

TGRNet: A Table Graph Reconstruction Network for Table Structure Recognition
Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, Qingyong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1295-1304

A table arranging data in rows and columns is a very effective data structure, which has been widely used in business and scientific research. Considering large-scale tabular data in online and offline documents, automatic table recognition has attracted increasing attention from the document analysis community. Though human can easily understand the structure of tables, it remains a challenge for machines to understand that, especially due to a variety of different table layouts and styles. Existing methods usually model a table as either the markup sequence or the adjacency matrix between different table cells, failing to address the importance of the logical location of table cells, e.g., a cell is located in the first row and the second column of the table. In this paper, we reformulate the problem of table structure recognition as the table graph reconstruction, and propose an end-to-end trainable table graph reconstruction network (TGRNet) for table structure recognition. Specifically, the proposed method has two main branches, a cell detection branch and a cell logical location branch, to jointly predict the spatial location and the logical location of different cells. Experimental results on three popular table recognition datasets and a new dataset with table graph annotations (TableGraph-350K) demonstrate the effectiveness of the proposed TGRNet for table structure recognition. Code and annotations will be made publicly available.

****

Mixture-Based Feature Space Learning for Few-Shot Image Classification
Arman Afrasiyabi, Jean-François Lalonde, Christian Gagné; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9041-9051

We introduce Mixture-based Feature Space Learning (MixtFSL) for obtaining a rich and robust feature representation in the context of few-shot image classification. Previous works have proposed to model each base class either with a single point or with a mixture model by relying on offline clustering algorithms. In contrast, we propose to model base classes with mixture models by simultaneously training the feature extractor and learning the mixture model parameters in an online manner. This results in a richer and more discriminative feature space which can be employed to classify novel examples from very few samples. Two main stages are proposed to train the MixtFSL model. First, the multimodal mixtures for each base class and the feature extractor parameters are learned using a combination of two loss functions. Second, the resulting network and mixture models are progressively refined through a leader-follower learning procedure, which uses the current estimate as a "target" network. This target network is used to make a consistent assignment of instances to mixture components, which increases performance and stabilizes training. The effectiveness of our end-to-end feature space learning approach is demonstrated with extensive experiments on four standard datasets and four backbones. Notably, we demonstrate that when we combine our robust representation with recent alignment-based approaches, we achieve new state-of-the-art results in the inductive setting, with an absolute accuracy for 5-shot classification of 82.45 on miniImageNet, 88.20 with tieredImageNet, and 60.70 in FC100 using the ResNet-12 backbone.
****

Learning a Sketch Tensor Space for Image Inpainting of Man-Made Scenes
Chenjie Cao, Yanwei Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14509-14518
This paper studies the task of inpainting man-made scenes. It is very challenging due to the difficulty in preserving the visual patterns of images, such as edges, lines, and junctions. Especially, most previous works are failed to restore the object/building structures for images of man-made scenes. To this end, this paper proposes learning a Sketch Tensor (ST) space for inpainting man-made scenes. Such a space is learned to restore the edges, lines, and junctions in images, and thus makes reliable predictions of the holistic image structures. To facilitate the structure refinement, we propose a Multi-scale Sketch Tensor inpainting (MST) network, with a novel encoder-decoder structure. The encoder extracts lines and edges from the input images to project them into an ST space. From this space, the decoder is learned to restore the input images. Extensive experiments validate the efficacy of our model. Furthermore, our model can also achieve competitive performance in inpainting general nature images over the competitors.
****

MicroNet: Improving Image Recognition With Extremely Low FLOPs
Yunsheng Li, Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Lu Yuan, Zicheng Liu, Lei Zhang, Nuno Vasconcelos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 468-477
This paper aims at addressing the problem of substantial performance degradation at extremely low computational cost (e.g. 5M FLOPs on ImageNet classification). We found that two factors, sparse connectivity and dynamic activation function, are effective to improve the accuracy. The former avoids the significant reduction of network width, while the latter mitigates the detriment of reduction in network depth. Technically, we propose micro-factorized convolution, which factorizes a convolution matrix into low rank matrices, to integrate sparse connectivity into convolution. We also present a new dynamic activation function, named Dynamic Shift Max, to improve the non-linearity via maxing out multiple dynamic fusions between an input feature map and its circular channel shift. Building upon these two new operators, we arrive at a family of networks, named MicroNet, that achieves significant performance gains over the state of the art in the low FLOP regime. For instance, under the constraint of 12M FLOPs, MicroNet achieves 59.4% top-1 accuracy on ImageNet classification, outperforming MobileNetV3 by 9.6%. Source code is at https://github.com/liyunsheng13/micronet.
****

Learning Canonical 3D Object Representation for Fine-Grained Recognition

Sunghun Joung, Seungryong Kim, Minsu Kim, Ig-Jae Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1035-1045

We propose a novel framework for fine-grained object recognition that learns to recover object variation in 3D space from a single image, trained on an image collection without using any ground-truth 3D annotation. We accomplish this by representing an object as a composition of 3D shape and its appearance, while eliminating the effect of camera viewpoint, in a canonical configuration. Unlike conventional methods modeling spatial variation in 2D images only, our method is capable of reconfiguring the appearance feature in a canonical 3D space, thus enabling the subsequent object classifier to be invariant under 3D geometric variation. Our representation also allows us to go beyond existing methods, by incorporating 3D shape variation as an additional cue for object recognition. To learn the model without ground-truth 3D annotation, we deploy a differentiable renderer in an analysis-by-synthesis framework. By incorporating 3D shape and appearance jointly in a deep representation, our method learns the discriminative representation of the object and achieves competitive performance on fine-grained image recognition and vehicle re-identification. We also demonstrate that the performance of 3D shape reconstruction is improved by learning fine-grained shape deformation in a boosting manner.

****

Multi-Class Multi-Instance Count Conditioned Adversarial Image Generation

Amrutha Saseendran, Kathrin Skubch, Margret Keuper; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6762-6771

Image generation has rapidly evolved in recent years. Modern architectures for adversarial training allow to generate even high resolution images with remarkable quality. At the same time, more and more effort is dedicated towards controlling the content of generated images. In this paper, we take one further step in this direction and propose a conditional generative adversarial network (GAN) that generates images with a defined number of objects from given classes. This entails two fundamental abilities (1) being able to generate high-quality images given a complex constraint and (2) being able to count object instances per class in a given image. Our proposed model modularly extends the successful StyleGAN2 architecture with a count-based conditioning as well as with a regression sub-network to count the number of generated objects per class during training. In experiments on three different datasets, we show that the proposed model learns to generate images according to the given multiple-class count condition even in the presence of complex backgrounds. In particular, we propose a new dataset, CityCount, which is derived from the Cityscapes street scenes dataset, to evaluate our approach in a challenging and practically relevant scenario. An implementation is available at https://github.com/boschresearch/MCCGAN.

****

Specialize and Fuse: Pyramidal Output Representation for Semantic Segmentation

Chi-Wei Hsiao, Cheng Sun, Hwann-Tzong Chen, Min Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7137-7146

We present a novel pyramidal output representation to ensure parsimony with our "specialize and fuse" process for semantic segmentation. A pyramidal "output" representation consists of coarse-to-fine levels, where each level is "specialize" in a different class distribution (e.g., more stuff than things classes at coarser levels). Two types of pyramidal outputs (i.e., unity and semantic pyramid) are "fused" into the final semantic output, where the unity pyramid indicates unity-cells (i.e., all pixels in such cell share the same semantic label). The process ensures parsimony by predicting a relatively small number of labels for unity-cells (e.g., a large cell of grass) to build the final semantic output. In addition to the "output" representation, we design a coarse-to-fine contextual module to aggregate the "features" representation from different levels. We validate the effectiveness of each key module in our method through comprehensive ablation studies. Finally, our approach achieves state-of-the-art performance on three widely-used semantic segmentation datasets---ADE20K, COCO-Stuff, and Pascal-Context.

****
DC-ShadowNet: Single-Image Hard and Soft Shadow Removal Using Unsupervised Domai n-Classifier Guided Network
Yeying Jin, Aashish Sharma, Robby T. Tan; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2021, pp. 5027-5036
Shadow removal from a single image is generally still an open problem. Most exis ting learning-based methods use supervised learning and require a large number o f paired images (shadow and corresponding non-shadow images) for training. A rec ent unsupervised method, Mask-ShadowGAN, addresses this limitation. However, it requires a binary mask to represent shadow regions, making it inapplicable to so ft shadows. To address the problem, in this paper, we propose an unsupervised do main-classifier guided shadow removal network, DC-ShadowNet. Specifically, we pr opose to integrate a shadow/shadow-free domain classifier into a generator and i ts discriminator, enabling them to focus on shadow regions. To train our network , we introduce novel losses based on physics-based shadow-free chromaticity, sha dow-robust perceptual features, and boundary smoothness. Moreover, we show that our network being unsupervised can be used for test-time training that further i mproves the results. Our experiments show that all these novel components allow our method to handle soft shadows, and also to perform better on hard shadows bo th quantitatively and qualitatively than the existing state-of-the-art shadow re moval methods.
****
Scalable Vision Transformers With Hierarchical Pooling
Zizheng Pan, Bohan Zhuang, Jing Liu, Haoyu He, Jianfei Cai; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 377-386
The recently proposed Visual image Transformers (ViT) with pure attention have a chieved promising performance on image recognition tasks, such as image classifi cation. However, the routine of the current ViT model is to maintain a full-leng th patch sequence during inference, which is redundant and lacks hierarchical re presentation. To this end, we propose a Hierarchical Visual Transformer (HVT) wh ich progressively pools visual tokens to shrink the sequence length and hence re duces the computational cost, analogous to the feature maps downsampling in Conv olutional Neural Networks (CNNs). It brings a great benefit that we can increase the model capacity by scaling dimensions of depth/width/resolution/patch size w ithout introducing extra computational complexity due to the reduced sequence le ngth. Moreover, we empirically find that the average pooled visual tokens contai n more discriminative information than the single class token. To demonstrate th e improved scalability of our HVT, we conduct extensive experiments on the image classification task. With comparable FLOPs, our HVT outperforms the competitive baselines on ImageNet and CIFAR-100 datasets. Code is available at https://gith ub.com/MonashAI/HVT.
****
Learning Instance-Level Spatial-Temporal Patterns for Person Re-Identification
Min Ren, Lingxiao He, Xingyu Liao, Wu Liu, Yunlong Wang, Tieniu Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1 4930-14939
Person re-identification (Re-ID) aims to match pedestrians under dis-joint camer as. Most Re-ID methods formulate it as visual representation learning and image search, and its accuracy is consequently affected greatly by the search space. S patial-temporal information has been proven to be efficient to filter irrelevant negative samples and significantly improve Re-ID accuracy. However, existing sp atial-temporal person Re-ID methods are still rough and do not exploit spatial-t emporal information sufficiently. In this paper, we propose a novel instance-lev el and spatial-temporal disentangled Re-ID method (InSTD), to improve Re-ID accu racy. In our proposed framework, personalized information such as moving directi on is explicitly considered to further narrow down the search space. Besides, th e spatial-temporal transferring probability is disentangled from joint distribut ion to marginal distribution, so that outliers can also be well modeled. Abundan t experimental analyses on two datasets are presented, which demonstrates the su periority and provides more insights into our method. The proposed method achiev

es mAP of 90.8% on Market-1501 and 89.1% on DukeMTMC-reID, improving from the b
aseline 82.2% and 72.7%, respectively. Besides, in order to provide a better ben
chmark for person re-identification, we release a cleaned data list of DukeMTMC-
reID with this paper: https://github.com/RenMin1991/cleaned-DukeMTMC-reID.
****

EgoRenderer: Rendering Human Avatars From Egocentric Camera Images
Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, Christian Theobalt; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
021, pp. 14528-14538
We present EgoRenderer, a system for rendering full-body neural avatars of a per
son captured by a wearable, egocentric fisheye camera that is mounted on a cap o
r a VR headset. Our system renders photorealistic novel views of the actor and h
er motion from arbitrary virtual camera locations. Rendering full-body avatars f
rom such egocentric images come with unique challenges due to the top-down view
and large distortions. We tackle these challenges by decomposing the rendering p
rocess into several steps, including texture synthesis, pose construction, and n
eural image translation. For texture synthesis, we propose Ego-DPNet, a neural n
etwork that infers dense correspondences between the input fisheye images and an
 underlying parametric body model, and to extract textures from egocentric input
s. In addition, to encode dynamic appearances, our approach also learns an impli
cit texture stack that captures detailed appearance variation across poses and v
iewpoints. For correct pose generation, we first estimate body pose from the ego
centric view using a parametric model. We then synthesize an external free-viewp
oint pose image by projecting the parametric model to the user-specified target
viewpoint. We next combine the target pose image and the textures into a combine
d feature image, which is transformed into the output color image using a neural
 image translation network. Experimental evaluations show that EgoRenderer is ca
pable of generating realistic free-viewpoint avatars of a person wearing an egoc
entric camera. Comparisons to several baselines demonstrate the advantages of ou
r approach.
****

Generative Adversarial Registration for Improved Conditional Deformable Template
s
Neel Dey, Mengwei Ren, Adrian V. Dalca, Guido Gerig; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 3929-3941
Deformable templates are essential to large-scale medical image registration, se
gmentation, and population analysis. Current conventional and deep network-based
 methods for template construction use only regularized registration objectives
and often yield templates with blurry and/or anatomically implausible appearance
, confounding downstream biomedical interpretation. We reformulate deformable re
gistration and conditional template estimation as an adversarial game wherein we
 encourage realism in the moved templates with a generative adversarial registra
tion framework conditioned on flexible image covariates. The resulting templates
 exhibit significant gain in specificity to attributes such as age and disease,
better fit underlying group-wise spatiotemporal trends, and achieve improved sha
rpness and centrality. These improvements enable more accurate population modeli
ng with diverse covariates for standardized downstream analyses and easier anato
mical delineation for structures of interest.
****

Visual Graph Memory With Unsupervised Representation for Visual Navigation
Obin Kwon, Nuri Kim, Yunho Choi, Hwiyeon Yoo, Jeongho Park, Songhwai Oh; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 15890-15899
We present a novel graph-structured memory for visual navigation, called visual
graph memory (VGM), which consists of unsupervised image representations obtaine
d from navigation history. The proposed VGM is constructed incrementally based o
n the similarities among the unsupervised representations of observed images, an
d these representations are learned from an unlabeled image dataset. We also pro
pose a navigation framework that can utilize the proposed VGM to tackle visual n
avigation problems. By incorporating a graph convolutional network and the atten

tion mechanism, the proposed agent refers to the VGM to navigate the environment while simultaneously building the VGM. Using the VGM, the agent can embed its navigation history and other useful task-related information. We validate our approach on the visual navigation tasks using the Habitat simulator with the Gibson dataset, which provides a photo-realistic simulation environment. The extensive experimental results show that the proposed navigation agent with VGM surpasses the state-of-the-art approaches on image-goal navigation tasks.

****

MGNet: Monocular Geometric Scene Understanding for Autonomous Driving
Markus Schön, Michael Buchholz, Klaus Dietmayer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15804-15815
We introduce MGNet, a multi-task framework for monocular geometric scene understanding. We define monocular geometric scene understanding as the combination of two known tasks: Panoptic segmentation and self-supervised monocular depth estimation. Panoptic segmentation captures the full scene not only semantically, but also on an instance basis. Self-supervised monocular depth estimation uses geometric constraints derived from the camera measurement model in order to measure depth from monocular video sequences only. To the best of our knowledge, we are the first to propose the combination of these two tasks in one single model. Our model is designed with focus on low latency to provide fast inference in real-time on a single consumer-grade GPU. During deployment, our model produces dense 3D point clouds with instance aware semantic labels from single high-resolution camera images. We evaluate our model on two popular autonomous driving benchmarks, i.e., Cityscapes and KITTI, and show competitive performance among other real-time capable methods. Source code is available at https://github.com/markusschoen/MGNet.

****

Auto-Parsing Network for Image Captioning and Visual Question Answering
Xu Yang, Chongyang Gao, Hanwang Zhang, Jianfei Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2197-2207
We propose an Auto-Parsing Network (APN) to discover and exploit the input data's hidden tree structures for improving the effectiveness of the Transformer-based vision-language systems. Specifically, we impose a Probabilistic Graphical Model (PGM) parameterized by the attention operations on each self-attention layer to incorporate sparse assumption. We use this PGM to softly segment an input sequence into a few clusters where each cluster can be treated as the parent of the inside entities. By stacking these PGM constrained self-attention layers, the clusters in a lower layer compose into a new sequence, and the PGM in a higher layer will further segment this sequence. Iteratively, a sparse tree can be implicitly parsed, and this tree's hierarchical knowledge is incorporated into the transformed embeddings, which can be used for solving the target vision-language tasks. Specifically, we showcase that our APN can strengthen Transformer based networks in two major vision-language tasks: Captioning and Visual Question Answering. Also, a PGM probability-based parsing algorithm is developed by which we can discover what the hidden structure of input is during the inference.

****

F-Drop&Match: GANs With a Dead Zone in the High-Frequency Domain
Shin'ya Yamaguchi, Sekitoshi Kanai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6743-6751
Generative adversarial networks built from deep convolutional neural networks (GANs) lack the ability to exactly replicate the high-frequency components of natural images. To alleviate this issue, we introduce two novel training techniques called frequency dropping (F-Drop) and frequency matching (F-Match). The key idea of F-Drop is to filter out unnecessary high-frequency components from the input images of the discriminators. This simple modification prevents the discriminators from being confused by perturbations of the high-frequency components. In addition, F-Drop makes the GANs focus on fitting in the low-frequency domain, in which there are the dominant components of natural images. F-Match minimizes the difference between real and fake images in the frequency domain for generating more realistic images. F-Match is implemented as a regularization term in the ob

jective functions of the generators; it penalizes the batch mean error in the fr
equency domain. F-Match helps the generators to fit in the high-frequency domain
 filtered out by F-Drop to the real image. We experimentally demonstrate that th
e combination of F-Drop and F-Match improves the generative performance of GANs
in both the frequency and spatial domain on multiple image benchmarks.
****

CryoDRGN2: Ab Initio Neural Reconstruction of 3D Protein Structures From Real Cr
yo-EM Images
Ellen D. Zhong, Adam Lerer, Joseph H. Davis, Bonnie Berger; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4066-4075
Protein structure determination from cryo-EM data requires reconstructing a 3D v
olume (or distribution of volumes) from many noisy and randomly oriented 2D proj
ection images. While the standard homogeneous reconstruction task aims to recove
r a single static structure, recently-proposed neural and non-neural methods can
 reconstruct distributions of structures, thereby enabling the study of protein
complexes that possess intrinsic structural or conformational heterogeneity. The
se heterogeneous reconstruction methods, however, require fixed image poses, whi
ch are typically estimated from an upstream homogeneous reconstruction and are n
ot guaranteed to be accurate under highly heterogeneous conditions. In this work
 we describe cryoDRGN2, an ab initio reconstruction algorithm, which can jointly
 estimate image poses and learn a neural model of a distribution of 3D structure
s on real heterogeneous cryo-EM data. To achieve this, we adapt search algorithm
s from the traditional cryo-EM literature, and describe the optimizations and de
sign choices required to make such a search procedure computationally tractable
in the neural model setting. We show that cryoDRGN2 is robust to the high noise
levels of real cryo-EM images, trains faster than earlier neural methods, and ac
hieves state-of-the-art performance on real cryo-EM datasets.
****

Generalized Shuffled Linear Regression
Feiran Li, Kent Fujiwara, Fumio Okura, Yasuyuki Matsushita; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6474-6483
We consider the shuffled linear regression problem where the correspondences bet
ween covariates and responses are unknown. While the existing formulation assume
s an ideal underlying bijection in which all pieces of data should match, such a
n assumption barely holds in real-world applications due to either missing data
or outliers. Therefore, in this work, we generalize the formulation of shuffled
linear regression to a broader range of conditions where only part of the data s
hould correspond. Moreover, we present a remarkably simple yet effective optimiz
ation algorithm with guaranteed global convergence. Distinct tasks validate the
effectiveness of the proposed method.
****

AESOP: Abstract Encoding of Stories, Objects, and Pictures
Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, Mubbasir Kapadia; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
, pp. 2052-2063
Visual storytelling and story comprehension are uniquely human skills that play
a central role in how we learn about and experience the world. Despite remarkabl
e progress in recent years in synthesis of visual and textual content in isolati
on and learning effective joint visual-linguistic representations, existing syst
ems still operate only at a superficial, factual level. With the goal of develop
ing systems that are able to comprehend rich human-generated narratives, and co-
create new stories, we introduce AESOP: a new dataset that captures the creative
 process associated with visual storytelling. Visual panels are composed of clip
-art objects with specific attributes enabling a broad range of creative express
ion. Using AESOP, we propose foundational storytelling tasks that are generative
 variants of story cloze tests, to better measure the creative and causal reason
ing ability required for visual storytelling. We further develop a generalized s
tory completion framework that models stories as the co-evolution of visual and
textual concepts. We benchmark the proposed approach with human baselines and ev
aluate using comprehensive qualitative and quantitative metrics. Our results hig

hlight key insights related to the dataset, modelling and evaluation of visual s
torytelling for future research in this promising field of study.
****

Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals
Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Luc Van Gool; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
, pp. 10052-10062
Being able to learn dense semantic representations of images without supervision
 is an important problem in computer vision. However, despite its significance,
this problem remains rather unexplored, with a few exceptions that considered un
supervised semantic segmentation on small-scale datasets with a narrow visual do
main. In this paper, we make a first attempt to tackle the problem on datasets t
hat have been traditionally utilized for the supervised case. To achieve this, w
e introduce a two-step framework that adopts a predetermined mid-level prior in
a contrastive optimization objective to learn pixel embeddings. This marks a lar
ge deviation from existing works that relied on proxy tasks or end-to-end cluste
ring. Additionally, we argue about the importance of having a prior that contain
s information about objects, or their parts, and discuss several possibilities t
o obtain such a prior in an unsupervised manner. Experimental evaluation shows t
hat our method comes with key advantages over existing works. First, the learned
 pixel embeddings can be directly clustered in semantic groups using K-Means on
PASCAL. Under the fully unsupervised setting, there is no precedent in solving t
he semantic segmentation task on such a challenging benchmark. Second, our repre
sentations can improve over strong baselines when transferred to new datasets, e
.g. COCO and DAVIS. The code is available.
****

Graph Contrastive Clustering
Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang N
ie, Zhouchen Lin, Xian-Sheng Hua; Proceedings of the IEEE/CVF International Conf
erence on Computer Vision (ICCV), 2021, pp. 9224-9233
Recently, some contrastive learning methods have been proposed to simultaneously
 learn representations and clustering assignments, achieving significant improve
ments. However, these methods do not take the category information and clusterin
g objective into consideration, thus the learned representations are not optimal
 for clustering and the performance might be limited. Towards this issue, we fir
st propose a novel graph contrastive learning framework, which is then applied t
o the clustering task and we come up with the Graph Constrastive Clustering (GCC
) method. Different from basic contrastive clustering that only assumes an image
 and its augmentation should share similar representation and clustering assignm
ents, we lift the instance-level consistency to the cluster-level consistency wi
th the assumption that samples in one cluster and their augmentations should all
 be similar. Specifically, on the one hand, the graph Laplacian based contrastiv
e loss is proposed to learn more discriminative and clustering-friendly features
. On the other hand, a novel graph-based contrastive learning strategy is propos
ed to learn more compact clustering assignments. Both of them incorporate the la
tent category information to reduce the intra-cluster variance while increasing
the inter-cluster variance. Experiments on six commonly used datasets demonstrat
e the superiority of our proposed approach over the state-of-the-art methods.
****

LFI-CAM: Learning Feature Importance for Better Visual Explanation
Kwang Hee Lee, Chaewon Park, Junghyun Oh, Nojun Kwak; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2021, pp. 1355-1363
Class Activation Mapping (CAM) is a powerful technique used to understand the de
cision making of Convolutional Neural Network (CNN) in computer vision. Recently
, there have been attempts not only to generate better visual explanations, but
also to improve classification performance using visual explanations. However, p
revious works still have their own drawbacks. In this paper, we propose a novel
architecture, LFI-CAM***(Learning Feature Importance Class Activation Mapping),
which is trainable for image classification and visual explanation in an end-to-
end manner. LFI-CAM generates attention map for visual explanation during forwar

d propagation, and simultaneously uses attention map to improve classification p
erformance through the attention mechanism. Feature Importance Network (FIN) foc
uses on learning the feature importance instead of directly learning the attenti
on map to obtain a more reliable and consistent attention map. We confirmed that
 LFI-CAM is optimized not only by learning the feature importance but also by en
hancing the backbone feature representation to focus more on important features
 of the input image. Experiments show that LFI-CAM outperforms baseline models' a
ccuracy on classification tasks as well as significantly improves on previous wo
rks in terms of attention map quality and stability over different hyper-paramet
ers.
****

InstanceRefer: Cooperative Holistic Understanding for Visual Grounding on Point Clouds Through Instance Multi-Level Contextual Referring
Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, Shuguang Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1791-1800
Compared with the visual grounding on 2D images, the natural-language-guided 3D
 object localization on point clouds is more challenging. In this paper, we propo
se a new model, named InstanceRefer, to achieve a superior 3D visual grounding t
hrough the grounding-by-matching strategy. In practice, our model first predicts
 the target category from the language descriptions using a simple language clas
sification model. Then based on the category, our model sifts out a small number
 of instance candidates (usually less than 20) from the panoptic segmentation on
 point clouds. Thus, the non-trivial 3D visual grounding task has been effective
ly re-formulated as a simplified instance-matching problem, considering that ins
tance-level candidates are more rational than the redundant 3D object proposals.
 Subsequently, for each candidate, we perform the multi-level contextual inferen
ce, i.e., referring from instance attribute perception, instance-to-instance rel
ation perception, and instance-to-background global localization perception, res
pectively. Eventually, the most relevant candidate is selected and localized by
 ranking confidence scores, which are obtained by the cooperative holistic visual
-language feature matching. Experiments confirm that our method outperforms prev
ious state-of-the-arts on ScanRefer online benchmark (ranked 1st place) and Nr3D
/Sr3D datasets.
****

Temporal-Wise Attention Spiking Neural Networks for Event Streams Classification
Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, Guo
qi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2021, pp. 10221-10230
How to effectively and efficiently deal with spatio-temporal event streams, wher
e the events are generally sparse and non-uniform and have the us temporal resol
ution, is of great value and has various real-life applications. Spiking neural
network (SNN), as one of the brain-inspired event-triggered computing models, ha
s the potential to extract effective spatio-temporal features from the event str
eams. However, when aggregating individual events into frames with a new higher
 temporal resolution, existing SNN models do not attach importance to that the se
rial frames have different signal-to-noise ratios since event streams are sparse
 and non-uniform. This situation interferes with the performance of existing SNN
s. In this work, we propose a temporal-wise attention SNN (TA-SNN) model to lear
n frame-based representation for processing event streams. Concretely, we extend
 the attention concept to temporal-wise input to judge the significance of frame
s for the final decision at the training stage, and discard the irrelevant frame
s at the inference stage. We demonstrate that TA-SNN models improve the accuracy
 of event streams classification tasks. We also study the impact of multiple-sca
le temporal resolutions for frame-based representation. Our approach is tested o
n three different classification tasks: gesture recognition, image classificatio
n, and spoken digit recognition. We report the state-of-the-art results on these
 tasks, and get the essential improvement of accuracy (almost 19%) for gesture r
ecognition with only 60 ms.
****

Encoder-Decoder With Multi-Level Attention for 3D Human Shape and Pose Estimation

Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13033-13042

3D human shape and pose estimation is the essential task for human motion analysis, which is widely used in many 3D applications. However, existing methods cannot simultaneously capture the relations at multiple levels, including spatial-temporal level and human joint level. Therefore they fail to make accurate predictions in some hard scenarios when there is cluttered background, occlusion, or extreme pose. To this end, we propose Multi-level Attention Encoder-Decoder Network (MAED), including a Spatial-Temporal Encoder (STE) and a Kinematic Topology Decoder (KTD) to model multi-level attentions in a unified framework. STE consists of a series of cascaded blocks based on Multi-Head Self-Attention, and each block uses two parallel branches to learn spatial and temporal attention respectively. Meanwhile, KTD aims at modeling the joint level attention. It regards pose estimation as a top-down hierarchical process similar to SMPL kinematic tree. With the training set of 3DPW, MAED outperforms previous state-of-the-art methods by 6.2, 7.2, and 2.4 mm of PA-MPJPE on the three widely used benchmarks 3DPW, MPI-INF-3DHP, and Human3.6M respectively. Our code is available at https://github.com/ziniuwan/maed.

****

Adaptive Hierarchical Graph Reasoning With Semantic Coherence for Video-and-Language Inference

Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, Yueting Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1867-1877

Video-and-Language Inference is a recently proposed task for joint video-and-language understanding. This new task requires a model to draw inference on whether a natural language statement entails or contradicts a given video clip. In this paper, we study how to address three critical challenges for this task: judging the global correctness of the statement involved multiple semantic meanings, joint reasoning over video and subtitles, and modeling long-range relationships and complex social interactions. First, we propose an adaptive hierarchical graph network that achieves in-depth understanding of the video over complex interactions. Specifically, it performs joint reasoning over video and subtitles in three hierarchies, where the graph structure is adaptively adjusted according to the semantic structures of the statement. Secondly, we introduce semantic coherence learning to explicitly encourage the semantic coherence of the adaptive hierarchical graph network from three hierarchies. The semantic coherence learning can further improve the alignment between vision and linguistics, and the coherence across a sequence of video segments. Experimental results show that our method significantly outperforms the baseline by a large margin.

****

Transductive Few-Shot Classification on the Oblique Manifold

Guodong Qi, Huimin Yu, Zhaohui Lu, Shuzhao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8412-8422

Few-shot learning (FSL) attempts to learn with limited data. In this work, we perform the feature extraction in the Euclidean space and the geodesic distance metric on the Oblique Manifold (OM). Specially, for better feature extraction, we propose a non-parametric Region Self-attention with Spatial Pyramid Pooling (RSSPP), which realizes a trade-off between the generalization and the discriminative ability of the single image feature. Then, we embed the feature to OM as a point. Furthermore, we design an Oblique Distance-based Classifier (ODC) that achieves classification in the tangent spaces which better approximate OM locally by learnable tangency points. Finally, we introduce a new method for parameters initialization and a novel loss function in the transductive settings. Extensive experiments demonstrate the effectiveness of our algorithm and it outperforms state-of-the-art methods on the popular benchmarks: mini-ImageNet, tiered-ImageNet, and Caltech-UCSD Birds-200-2011 (CUB).

****
iNAS: Integral NAS for Device-Aware Salient Object Detection
Yu-Chao Gu, Shang-Hua Gao, Xu-Sheng Cao, Peng Du, Shao-Ping Lu, Ming-Ming Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4934-4944
Existing salient object detection (SOD) models usually focus on either backbone feature extractors or saliency heads, ignoring their relations. A powerful backbone could still achieve sub-optimal performance with a weak saliency head and vice versa. Moreover, the balance between model performance and inference latency poses a great challenge to model design, especially when considering different deployment scenarios. Considering all components in an integral neural architecture search (iNAS) space, we propose a flexible device-aware search scheme that only trains the SOD model once and quickly finds high-performance but low-latency models on multiple devices. An evolution search with latency-group sampling (LGS) is proposed to explore the entire latency area of our enlarged search space. Models searched by iNAS achieve similar performance with SOTA methods but reduce the 3.8x, 3.3x, 2.6x, 1.9x latency on Huawei Nova6 SE, Intel Core CPU, the Jetson Nano, and Nvidia Titan Xp. The code is released at https://mmcheng.net/inas/.
****
Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection
Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, Chunjing Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2723-2732
We present a flexible and high-performance framework, named Pyramid R-CNN, for two-stage 3D object detection from point clouds. Current approaches generally rely on the points or voxels of interest for RoI feature extraction on the second stage, but cannot effectively handle the sparsity and non-uniform distribution of those points, and this may result in failures in detecting objects that are far away. To resolve the problems, we propose a novel second-stage module, named pyramid RoI head, to adaptively learn the features from the sparse points of interest. The pyramid RoI head consists of three key components. Firstly, we propose the RoI-grid Pyramid, which addresses the sparsity problem by extensively collecting points of interest for each RoI in a pyramid manner. Secondly, we propose RoI-grid Attention, a new operation that can encode richer information from sparse points by incorporating conventional attention-based and graph-based point operators into a unified formulation. Thirdly, we propose the Density-Aware Radius Prediction (DARP) module, which can adapt to different point density levels by dynamically adjusting the focusing range of RoIs. Combining the three components, our pyramid RoI head is robust to the sparse and imbalanced circumstances, and can be applied upon various 3D backbones to consistently boost the detection performance. Extensive experiments show that Pyramid R-CNN outperforms the state-of-the-art 3D detection models by a large margin on both the KITTI dataset and the Waymo Open dataset.
****
Graph-BAS3Net: Boundary-Aware Semi-Supervised Segmentation Network With Bilateral Graph Convolution
Huimin Huang, Lanfen Lin, Yue Zhang, Yingying Xu, Jing Zheng, XiongWei Mao, Xiaohan Qian, Zhiyi Peng, Jianying Zhou, Yen-Wei Chen, Ruofeng Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7386-7395
Semi-supervised learning (SSL) algorithms have attracted much attentions in medical image segmentation by leveraging unlabeled data, which challenge in acquiring massive pixel-wise annotated samples. However, most of the existing SSLs neglected the geometric shape constraint in object, leading to unsatisfactory boundary and non-smooth of object. In this paper, we propose a novel boundary-aware semi-supervised medical image segmentation network, named Graph-BAS3Net, which incorporates the boundary information and learns duality constraints between semantics and geometrics in the graph domain. Specifically, the proposed method consists of two components: a multi-task learning framework BAS3Net and a graph-based c

ross-task module BGCM. The BAS3Net improves the existing GAN-based SSL by adding a boundary detection task, which encodes richer features of object shape and surface. Moreover, the BGCM further explores the co-occurrence relations between the semantics segmentation and boundary detection task, so that the network learns stronger semantic and geometric correspondences from both labeled and unlabeled data. Experimental results on the LiTS dataset and COVID-19 dataset confirm that our proposed Graph-BAS3 Net outperforms the state-of-the-art methods in semi-supervised segmentation task.

****

The Animation Transformer: Visual Correspondence via Segment Matching

Evan Casey, Víctor Pérez, Zhuoru Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11323-11332

Visual correspondence is a fundamental building block on the way to building assistive tools for hand-drawn animation. However, while a large body of work has focused on learning visual correspondences at the pixel-level, few approaches have emerged to learn correspondence at the level of line enclosures (segments) that naturally occur in hand-drawn animation. Exploiting this structure in animation has numerous benefits: it avoids the memory complexity of pixel attention over high resolution images and enables the use of real-world animation datasets that contain correspondence information at the level of per-segment colors. To that end, we propose the Animation Transformer (AnT) which uses a Transformer-based architecture to learn the spatial and visual relationships between segments across a sequence of images. By leveraging a forward match loss and a cycle consistency loss our approach attains excellent results compared to state-of-the-art pixel approaches on challenging datasets from real animation productions that lack ground-truth correspondence labels.

****

CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification

Chun-Fu (Richard) Chen, Quanfu Fan, Rameswar Panda; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 357-366

The recently developed vision transformer (ViT) has achieved promising results on image classification compared to convolutional neural networks. Inspired by this, in this paper, we study how to learn multi-scale feature representations in transformer models for image classification. To this end, we propose a dual-branch transformer to combine image patches (i.e., tokens in a transformer) of different sizes to produce stronger image features. Our approach processes small-patch and large-patch tokens with two separate branches of different computational complexity and these tokens are then fused purely by attention multiple times to complement each other. Furthermore, to reduce computation, we develop a simple yet effective token fusion module based on cross attention, which uses a single token for each branch as a query to exchange information with other branches. Our proposed cross-attention only requires linear time for both computational and memory complexity instead of quadratic time otherwise. Extensive experiments demonstrate that our approach performs better than or on par with several concurrent works on vision transformer, in addition to efficient CNN models. For example, on the ImageNet1K dataset, with some architectural changes, our approach outperforms the recent DeiT by a large margin of 2% with a small to moderate increase in FLOPs and model parameters. Our source codes and models are available at https://github.com/IBM/CrossViT.

****

Weak Adaptation Learning: Addressing Cross-Domain Data Insufficiency With Weak Annotator

Shichao Xu, Lixu Wang, Yixuan Wang, Qi Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8917-8926

Data quantity and quality are crucial factors for data-driven learning methods. In some target problem domains, there are not many data samples available, which could significantly hinder the learning process. While data from similar domains may be leveraged to help through domain adaptation, obtaining high-quality labeled data for those source domains themselves could be difficult or costly. To a

ddress such challenges on data insufficiency for classification problem in a tar
get domain, we propose a weak adaptation learning (WAL) approach that leverages
unlabeled data from a similar source domain, a low-cost weak annotator that prod
uces labels based on task-specific heuristics, labeling rules, or other methods
(albeit with inaccuracy), and a small amount of labeled data in the target domai
n. Our approach first conducts a theoretical analysis on the error bound of the
trained classifier with respect to the data quantity and the performance of the
weak annotator, and then introduces a multi-stage weak adaptation learning metho
d to learn an accurate classifier by lowering the error bound. Our experiments d
emonstrate the effectiveness of our approach in learning an accurate classifier
with limited labeled data in the target domain and unlabeled data in the source
domain.
****

Building-GAN: Graph-Conditioned Architectural Volumetric Design Generation
Kai-Hung Chang, Chin-Yi Cheng, Jieliang Luo, Shingo Murata, Mehdi Nourbakhsh, Yo
shito Tsuji; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2021, pp. 11956-11965
Volumetric design is the first and critical step for professional building desig
n, where architects not only depict the rough 3D geometry of the building but al
so specify the programs to form a 2D layout on each floor. Though 2D layout gene
ration for a single story has been widely studied, there is no developed method
for multi-story buildings. This paper focuses on volumetric design generation co
nditioned on an input program graph. Instead of outputting dense 3D voxels, we p
ropose a new 3D representation named voxel graph that is both compact and expres
sive for building geometries. Our generator is a cross-modal graph neural networ
k that uses a pointer mechanism to connect the input program graph and the outpu
t voxel graph, and the whole pipeline is trained using the adversarial framework
. The generated designs are evaluated qualitatively by a user study and quantita
tively using three metrics: quality, diversity, and connectivity accuracy. We sh
ow that our model generates realistic 3D volumetric designs and outperforms prev
ious methods and baselines.
****

Scribble-Supervised Semantic Segmentation Inference
Jingshan Xu, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yuge Huang, Pengcheng Shen, Sh
aoxin Li, Jian Yang; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2021, pp. 15354-15363
In this paper, we propose a progressive segmentation inference (PSI) framework t
o tackle with scribble-supervised semantic segmentation. In virtue of latent con
textual dependency, we encapsulate two crucial cues, contextual pattern propagat
ion and semantic label diffusion, to enhance and refine pixel-level segmentation
 results from partially known seeds. In contextual pattern propagation, differen
t-granular contextual patterns are correlated and leveraged to properly diffuse
pattern information based on graphical model, so as to increase the inference co
nfidence of pixel label prediction. Further, depending on high confidence scores
 of estimated pixels, the initial annotated seeds are progressively spread over
the image through dynamically learning an adaptive decision strategy. The two cu
es are finally modularized to form a close-looping update process during pixel-w
ise label inference. Extensive experiments demonstrate that our proposed progres
sive segmentation inference can benefit from the combination of spatial and sema
ntic context cues, and meantime achieve the state-of-the-art performance on two
public scribble segmentation datasets.
****

Improve Unsupervised Pretraining for Few-Label Transfer
Suichan Li, Dongdong Chen, Yinpeng Chen, Lu Yuan, Lei Zhang, Qi Chu, Bin Liu, Ne
nghai Yu; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2021, pp. 10201-10210
Unsupervised pretraining has achieved great success and many recently works have
 shown unsupervised pretraining can achieve comparable or even slightly better t
ransfer performance than supervised pretraining on downstream target datasets. B
ut in this paper, we find this conclusion may not hold when the target dataset h

as very few labeled samples for finetuning, ie, few-label transfer. We analyze t
he possible reason from the clustering perspective: 1) The clustering quality of
 target samples is of great importance to few-label transfer; 2) Though contrast
ive learning is essentially to learn how to cluster, its clustering quality is s
till inferior to supervised pretraining due to lack of label supervision. Based
on the analysis, we interestingly discover that only involving some unlabeled ta
rget domain into the unsupervised pretraining can improve the clustering quality
, subsequently reducing the transfer performance gap with supervised pretraining
. This finding also motivates us to propose a new progressive few-label transfer
 algorithm for real applications, which aims to maximize the transfer performanc
e under a limited annotation budget. To support our analysis and proposed method
, we conduct extensive experiments on nine different target datasets. Experiment
al results show our proposed method can significantly boost the few-label transf
er performance of unsupervised pretraining.
****

Image Inpainting via Conditional Texture and Structure Dual Generation
Xiefan Guo, Hongyu Yang, Di Huang; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2021, pp. 14134-14143
Deep generative approaches have recently made considerable progress in image inp
ainting by introducing structure priors. Due to the lack of proper interaction w
ith image texture during structure reconstruction, however, current solutions ar
e incompetent in handling the cases with large corruptions, and they generally s
uffer from distorted results. In this paper, we propose a novel two-stream netwo
rk for image inpainting, which models the structure-constrained texture synthesi
s and texture-guided structure reconstruction in a coupled manner so that they b
etter leverage each other for more plausible generation. Furthermore, to enhance
 the global consistency, a Bi-directional Gated Feature Fusion (Bi-GFF) module i
s designed to exchange and combine the structure and texture information and a C
ontextual Feature Aggregation (CFA) module is developed to refine the generated
contents by region affinity learning and multi-scale feature aggregation. Qualit
ative and quantitative experiments on the CelebA, Paris StreetView and Places2 d
atasets demonstrate the superiority of the proposed method. Our code is availabl
e at https://github.com/Xiefan-Guo/CTSDG.
****

Geometry-Aware Self-Training for Unsupervised Domain Adaptation on Object Point
Clouds
Longkun Zou, Hui Tang, Ke Chen, Kui Jia; Proceedings of the IEEE/CVF Internation
al Conference on Computer Vision (ICCV), 2021, pp. 6403-6412
The point cloud representation of an object can have a large geometric variation
 in view of inconsistent data acquisition procedure, which thus leads to domain
discrepancy due to diverse and uncontrollable shape representation cross dataset
s. To improve discrimination on unseen distribution of point-based geometries in
 a practical and feasible perspective, this paper proposes a new method of geome
try-aware self-training (GAST) for unsupervised domain adaptation of object poin
t cloud classification. Specifically, this paper aims to learn a domain-shared r
epresentation of semantic categories, via two novel self-supervised geometric le
arning tasks as feature regularization. On one hand, the representation learning
 is empowered by a linear mixup of point cloud samples with their self-generated
 rotation labels, to capture a global topological configuration of local geometr
ies. On the other hand, a diverse point distribution across datasets can be norm
alized with a novel curvature-aware distortion localization. Experiments on the
PointDA-10 dataset show that our GAST method can significantly outperform the st
ate-of-the-art methods.
****

Robustness and Generalization via Generative Adversarial Training
Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, Ser-Nam Lim; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 15711-15720
While deep neural networks have achieved remarkable success in various computer
vision tasks, they often fail to generalize to subtle variations of input images

. Several defenses have been proposed to improve the robustness against these va
riations. However, current defenses can only withstand the specific attack used
in training, and the models often remain vulnerable to other input variations. M
oreover, these methods often degrade performance of the model on clean images. I
n this paper, we present Generative Adversarial Training, an approach to simulta
neously improve the model's generalization and robustness to unseen adversarial
attacks. Instead of altering a single pre-defined aspect of images, we generate
a spectrum of low-level, mid-level and high-level changes using generative model
s with a disentangled latent space. Adversarial training with these examples ena
ble the model to withstand a wide range of attacks by observing a variety of inp
ut alterations during training. We show that our approach not only improves perf
ormance of the model on clean images but also makes it robust against unforeseen
 attacks and outperforms prior work. We validate effectiveness of our method by
demonstrating results on various tasks such as classification, semantic segmenta
tion and object detection.
****

Exploring Inter-Channel Correlation for Diversity-Preserved Knowledge Distillati
on
Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, Xiaodan
Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2021, pp. 8271-8280
Knowledge Distillation has shown very promising ability in transferring learned
representation from the larger model (teacher) to the smaller one (student). Des
pite many efforts, prior methods ignore the important role of retaining inter-ch
annel correlation of features, leading to the lack of capturing intrinsic distri
bution of the feature space and sufficient diversity properties of features in t
he teacher network. To solve the issue, we propose the novel Inter-Channel Corre
lation for Knowledge Distillation (ICKD), with which the diversity and homology
of the feature space of the student network can align with that of the teacher n
etwork. The correlation between these two channels is interpreted as diversity i
f they are irrelevant to each other, otherwise homology. Then the student is req
uired to mimic the correlation within its own embedding space. In addition, we i
ntroduce the grid-level inter-channel correlation, making it capable of dense pr
ediction tasks. Extensive experiments on two vision tasks, including ImageNet cl
assification and Pascal VOC segmentation, demonstrate the superiority of our ICK
D, which consistently outperforms many existing methods, advancing the state-of-
the-art in the fields of Knowledge Distillation. To our knowledge, we are the fi
rst method based on knowledge distillation boosts ResNet18 beyond 72% Top-1 accu
racy on ImageNet classification. Code is available at: https://github.com/ADLab-
AutoDrive/ICKD.
****

Class-Incremental Learning for Action Recognition in Videos
Jaeyoo Park, Minsoo Kang, Bohyung Han; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2021, pp. 13698-13707
We tackle catastrophic forgetting problem in the context of class-incremental le
arning for video recognition, which has not been explored actively despite the p
opularity of continual learning. Our framework addresses this challenging task b
y introducing time-channel importance maps and exploiting the importance maps fo
r learning the representations of incoming examples via knowledge distillation.
We also incorporate a regularization scheme in our objective function, which enc
ourages individual features obtained from different time steps in a video to be
uncorrelated and eventually improves accuracy by alleviating catastrophic forget
ting. We evaluate the proposed approach on brand-new splits of class-incremental
 action recognition benchmarks constructed upon the UCF101, HMDB51, and Somethin
g-Something V2 datasets, and demonstrate the effectiveness of our algorithm in c
omparison to the existing continual learning methods that are originally designe
d for image data.
****

Procrustean Training for Imbalanced Deep Learning
Han-Jia Ye, De-Chuan Zhan, Wei-Lun Chao; Proceedings of the IEEE/CVF Internation

al Conference on Computer Vision (ICCV), 2021, pp. 92-102
Neural networks trained with class-imbalanced data are known to perform poorly o n minor classes of scarce training data. Several recent works attribute this to over-fitting to minor classes. In this paper, we provide a novel explanation of this issue. We found that a neural network tends to first under-fit the minor cl asses by classifying most of their data into the major classes in early training epochs. To correct these wrong predictions, the neural network then must focus on pushing features of minor class data across the decision boundaries between m ajor and minor classes, leading to much larger gradients for features of minor c lasses. We argue that such an under-fitting phase over-emphasizes the competitio n between major and minor classes, hinders the neural network from learning the discriminative knowledge that can be generalized to test data, and eventually re sults in over-fitting. To address this issue, we propose a novel learning strate gy to equalize the training progress across classes. We mix features of the majo r class data with those of other data in a mini-batch, intentionally weakening t heir features to prevent a neural network from fitting them first. We show that this strategy can largely balance the training accuracy and feature gradients ac ross classes, effectively mitigating the under-fitting then over-fitting problem for minor class data. On several benchmark datasets, our approach achieves the state-of-the-art accuracy, especially for the challenging step-imbalanced cases.
****

Dynamic Network Quantization for Efficient Video Inference
Ximeng Sun, Rameswar Panda, Chun-Fu (Richard) Chen, Aude Oliva, Rogerio Feris, K ate Saenko; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2021, pp. 7375-7385
Deep convolutional networks have recently achieved great success in video recogn ition, yet their practical realization remains a challenge due to the large amou nt of computational resources required to achieve robust recognition. Motivated by the effectiveness of quantization for boosting efficiency, in this paper, we propose a dynamic network quantization framework, that selects optimal precision for each frame conditioned on the input for efficient video recognition. Specif ically, given a video clip, we train a very lightweight network in parallel with the recognition network, to produce a dynamic policy indicating which numerical precision to be used per frame in recognizing videos. We train both networks ef fectively using standard backpropagation with a loss to achieve both competitive performance and resource efficiency required for video recognition. Extensive e xperiments on four challenging diverse benchmark datasets demonstrate that our p roposed approach provides significant savings in computation and memory usage wh ile outperforming the existing state-of-the-art methods. Project page: https://c s-people.bu.edu/sunxm/VideoIQ/project.html.
****

Space-Time Crop & Attend: Improving Cross-Modal Video Representation Learning
Mandela Patrick, Po-Yao Huang, Ishan Misra, Florian Metze, Andrea Vedaldi, Yuki M. Asano, João F. Henriques; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2021, pp. 10560-10572
The quality of the image representations obtained from self-supervised learning depends strongly on the type of data augmentations used in the learning formulat ion. Recent papers have ported these methods from still images to videos and fou nd that leveraging both audio and video signals yields strong gains; however, th ey did not find that spatial augmentations such as cropping, which are very impo rtant for still images, work as well for videos. In this paper, we improve these formulations in two ways unique to the spatio-temporal aspect of videos. First, for space, we show that spatial augmentations such as cropping do work well for videos too, but that previous implementations, due to the high processing and m emory cost, could not do this at a scale sufficient for it to work well. To addr ess this issue, we first introduce Feature Crop, a method to simulate such augme ntations much more efficiently directly in feature space. Second, we show that a s opposed to naive average pooling, the use of transformer-based attention impro ves performance significantly, and is well suited for processing feature crops. Combining both of our discoveries into a new method, Space-time Crop & Attend (S

TiCA) we achieve state-of-the-art performance across multiple video-representation learning benchmarks. In particular, we achieve new state-of-the-art accuracies of 67.0% on HMDB-51 and 93.1% on UCF-101 when pre-training on Kinetics-400.
****

RDA: Robust Domain Adaptation via Fourier Adversarial Attacking
Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8988-8999
Unsupervised domain adaptation (UDA) involves a supervised loss in a labeled source domain and an unsupervised loss in an unlabeled target domain, which often faces more severe overfitting (than classical supervised learning) as the supervised source loss has clear domain gap and the unsupervised target loss is often noisy due to the lack of annotations. This paper presents RDA, a robust domain adaptation technique that introduces adversarial attacking to mitigate overfitting in UDA. We achieve robust domain adaptation by a novel Fourier adversarial attacking (FAA) method that allows large magnitude of perturbation noises but has minimal modification of image semantics, the former is critical to the effectiveness of its generated adversarial samples due to the existence of domain gaps. Specifically, FAA decomposes images into multiple frequency components (FCs) and generates adversarial samples by just perturbating certain FCs that capture little semantic information. With FAA-generated samples, the training can continue the random walk and drift into an area with a flat loss landscape, leading to more robust domain adaptation. Extensive experiments over multiple domain adaptation tasks show that RDA can work with different computer vision tasks with superior performance.
****

WB-DETR: Transformer-Based Detector Without Backbone
Fanfan Liu, Haoran Wei, Wenzhe Zhao, Guozhen Li, Jingquan Peng, Zihao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2979-2987
Transformer-based detector is a new paradigm in object detection, which aims to achieve pretty-well performance while eliminates the priori knowledge driven components, e.g., anchors, proposals and the NMS. DETR, the state-of-the-art model among them, is composed of three sub-modules, i.e., a CNN-based backbone and paired transformer encoder-decoder. The CNN is applied to extract local features and the transformer is used to capture global contexts. This pipeline, however, is not concise enough. In this paper, we propose WB-DETR (DETR-based detector Without Backbone) to prove that the reliance on CNN features extraction for a transformer-based detector is not necessary. Unlike the original DETR, WB-DETR is composed of only an encoder and a decoder without CNN backbone. For an input image, WB-DETR serializes it directly to encode the local features into each individual token. To make up the deficiency of transformer in modeling local information, we design an LIE-T2T (local information enhancement tokens to token) module to enhance the internal information of tokens after unfolding. Experimental results demonstrate that WB-DETR, the first pure-transformer detector without CNN to our knowledge, yields on par accuracy and faster inference speed with only half number of parameters compared with DETR baseline.
****

Worldsheet: Wrapping the World in a 3D Sheet for View Synthesis From a Single Image
Ronghang Hu, Nikhila Ravi, Alexander C. Berg, Deepak Pathak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12528-12537
We present Worldsheet, a method for novel view synthesis using just a single RGB image as input. The main insight is that simply shrink-wrapping a planar mesh sheet onto the input image, consistent with the learned intermediate depth, captures underlying geometry sufficient to generate photorealistic unseen views with large viewpoint changes. To operationalize this, we propose a novel differentiable texture sampler that allows our wrapped mesh sheet to be textured and rendered differentiably into an image from a target viewpoint. Our approach is category-agnostic, end-to-end trainable without using any 3D supervision, and requires a

single image at test time. We also explore a simple extension by stacking multiple layers of Worldsheets to better handle occlusions. Worldsheet consistently outperforms prior state-of-the-art methods on single-image view synthesis across several datasets. Furthermore, this simple idea captures novel views surprisingly well on a wide range of high-resolution in-the-wild images, converting them into navigable 3D pop-ups. Video results and code are available at https://worldsheet.github.io.

****

Patch2CAD: Patchwise Embedding Learning for In-the-Wild Shape Retrieval From a Single Image

Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, Angela Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12589-12599

3D perception of object shapes from RGB image input is fundamental towards semantic scene understanding, grounding image-based perception in our spatially 3-dimensional real-world environments. To achieve a mapping between image views of objects and 3D shapes, we leverage CAD model priors from existing large-scale databases, and propose a novel approach towards constructing a joint embedding space between 2D images and 3D CAD models in a patch-wise fashion -- establishing correspondences between patches of an image view of an object and patches of CAD geometry. This enables part similarity reasoning for retrieving similar CADs to a new image view without exact matches in the database. Our patch embedding provides more robust CAD retrieval for shape estimation in our end-to-end estimation of CAD model shape and pose for detected objects in a single input image. Experiments on in-the-wild, complex imagery from ScanNet show that our approach is more robust than state of the art in real-world scenarios without any exact CAD matches.

****

Perceptual Variousness Motion Deblurring With Light Global Context Refinement

Jichun Li, Weimin Tan, Bo Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4116-4125

Deep learning algorithms have made significant progress in dynamic scene deblurring. However, several challenges are still unsettled: 1) The degree and scale of blur in different regions of a blurred image can have a considerable variation in a large range. However, the traditional input pyramid or downscaling-upscaling, is designed to have limited and inflexible perceptual variousness to cope with large blur scale variation. 2) The nonlocal block is proved to be effective in the image enhancement tasks, but it requires high computation and memory cost. In this paper, we are the first to propose a light-weight globally-analyzing module into the image deblurring field, named Light Global Context Refinement (LGCR) module. With exponentially lower cost, it achieves even better performance than the nonlocal unit. Moreover, we propose the Perceptual Variousness Block (PVB) and PVB-piling strategy. By placing PVB repeatedly, the whole method possesses abundant reception field spectrum to be aware of the blur with various degrees and scales. Comprehensive experimental results from the different benchmarks and assessment metrics show that our method achieves excellent performance to set a new state-of-the-art in motion deblurring.

****

Self-Calibrating Neural Radiance Fields

Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, Jaesik Park; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5846-5854

In this work, we propose a camera self-calibration algorithm for generic cameras with arbitrary non-linear distortions. We jointly learn the geometry of the scene and the accurate camera parameters without any calibration objects. Our camera model consists of a pinhole model, a fourth order radial distortion, and a generic noise model that can learn arbitrary non-linear camera distortions. While traditional self-calibration algorithms mostly rely on geometric constraints, we additionally incorporate photometric consistency. This requires learning the geometry of the scene, and we use Neural Radiance Fields (NeRF). We also propose a new geometric loss function, viz., projected ray distance loss, to incorporate g

eometric consistency for complex non-linear camera models. We validate our appro ach on standard real image datasets and demonstrate that our model can learn the camera intrinsics and extrinsics (pose) from scratch without COLMAP initializat ion. Also, we show that learning accurate camera models in a differentiable mann er allows us to improve PSNR over baselines.
****

Motion Adaptive Pose Estimation From Compressed Videos
Zhipeng Fan, Jun Liu, Yao Wang; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2021, pp. 11719-11728
Human pose estimation from videos has many real-world applications. Existing met hods focus on applying models with a uniform computation profile on fully de- co ded frames, ignoring the freely available motion signals and motion-compensation residuals from the compressed stream. A novel model, called Motion Adaptive Pos e Net is proposed to exploit the compressed streams to efficiently decode pose s equences from videos. The model incorporates a Motion Compensated ConvLSTM to pr opagate the spatially aligned features, along with an adaptive gate to dynamical ly determine if the computationally expensive features should be extracted from fully decoded frames to compensate the motion-warped features, solely based on t he residual errors. Leveraging the informative yet readily available signals fro m compressed streams, we propagate the latent features through our Motion Adapti ve Pose Net efficiently. Our model outperforms the state-of-the-art models in po se- estimation accuracy on two widely used datasets with only around half of the computation complexity.
****

Learning Motion Priors for 4D Human Body Capture in 3D Scenes
Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, Siyu Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1134 3-11353
Recovering high-quality 3D human motion in complex scenes from monocular videos is important for many applications, ranging from AR/VR to robotics. However, cap turing realistic human-scene interactions, while dealing with occlusions and par tial views, is challenging; current approaches are still far from achieving comp elling results. We address this problem by proposing LEMO: LEarning human MOtion priors for 4D human body capture. By leveraging the large-scale motion capture dataset AMASS, we introduce a novel motion smoothness prior, which strongly redu ces the jitters exhibited by poses recovered over a sequence. Furthermore, to ha ndle contacts and occlusions occurring frequently in body-scene interactions, we design a contact friction term and a contact-aware motion infiller obtained via per-instance self-supervised training. To prove the effectiveness of the propos ed motion priors, we combine them into a novel pipeline for 4D human body captur e in 3D scenes. With our pipeline, we demonstrate high-quality 4D human body cap ture, reconstructing smooth motions and physically plausible body-scene interact ions. The code and data are available at https://sanweiliti.github.io/LEMO/LEMO. html.
****

Robust Automatic Monocular Vehicle Speed Estimation for Traffic Surveillance
Jerome Revaud, Martin Humenberger; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2021, pp. 4551-4561
Even though CCTV cameras are widely deployed for traffic surveillance and have t herefore the potential of becoming cheap automated sensors for traffic speed ana lysis, their large-scale usage toward this goal has not been reported yet. A key difficulty lies in fact in the camera calibration phase. Existing state-of-the- art methods perform the calibration using image processing or keypoint detection techniques that require high-quality video streams, yet typical CCTV footage is low-resolution and noisy. As a result, these methods largely fail in real-world conditions. In contrast, we propose two novel calibration techniques whose only inputs come from an off-the-shelf object detector. Both methods consider multip le detections jointly, leveraging the fact that cars have similar and well-known 3D shapes with normalized dimensions. The first one is based on minimizing an e nergy function corresponding to a 3D reprojection error, the second one instead

learns from synthetic training data to predict the scene geometry directly. Noti
cing the lack of speed estimation benchmarks faithfully reflecting the actual qu
ality of surveillance cameras, we introduce a novel dataset collected from publi
c CCTV streams. Experimental results conducted on three diverse benchmarks demon
strate excellent speed estimation accuracy that could enable the wide use of CCT
V cameras for traffic analysis, even in challenging conditions where state-of-th
e-art methods completely fail. Additional information can be found on our projec
t web page: https://rebrand.ly/nle-cctv
****

Face Image Retrieval With Attribute Manipulation
Alireza Zaeemzadeh, Shabnam Ghadar, Baldo Faieta, Zhe Lin, Nazanin Rahnavard, Mu
barak Shah, Ratheesh Kalarot; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2021, pp. 12116-12125
Current face image retrieval solutions are limited, since they treat different f
acial attributes the same and cannot incorporate user's preference for a subset
of attributes in their search criteria. This paper introduces a new face image r
etrieval framework, where the input face query is augmented by both an adjustmen
t vector that specifies the desired modifications to the facial attributes, and
a preference vector that assigns different levels of importance to different att
ributes. For example, a user can ask for retrieving images similar to a query im
age, but with a different hair color, and no preference for absence/presence of
eyeglasses in the results. To achieve this, we propose to disentangle the semant
ics, corresponding to various attributes, by learning a set of sparse and orthog
onal basis vectors in the latent space of StyleGAN. Such basis vectors are then
employed to decompose the dissimilarity between face images in terms of dissimil
arity between their attributes, assign preference to the attributes, and adjust
the attributes in the query. Enforcing sparsity on the basis vectors helps us to
 disentangle the latent space and adjust each attribute independently from other
 attributes, while enforcing orthogonality facilitates preference assignment and
 the dissimilarity decomposition. The effectiveness of our approach is illustrat
ed by achieving state-of-the-art results for the face image retrieval task.
****

RFNet: Region-Aware Fusion Network for Incomplete Multi-Modal Brain Tumor Segmen
tation
Yuhang Ding, Xin Yu, Yi Yang; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2021, pp. 3975-3984
Most existing brain tumor segmentation methods usually exploit multi-modal magne
tic resonance imaging (MRI) images to achieve high segmentation performance. How
ever, the problem of missing certain modality images often happens in clinical p
ractice, thus leading to severe segmentation performance degradation. In this wo
rk, we propose a Region-aware Fusion Network (RFNet) that is able to exploit dif
ferent combinations of multi-modal data adaptively and effectively for tumor seg
mentation. Considering different modalities are sensitive to different brain tum
or regions, we design a Region-aware Fusion Module (RFM) in RFNet to conduct mod
al feature fusion from available image modalities according to disparate regions
. Benefiting from RFM, RFNet can adaptively segment tumor regions from an incomp
lete set of multi-modal images by effectively aggregating modal features. Furthe
rmore, we also develop a segmentation-based regularizer to prevent RFNet from th
e insufficient and unbalanced training caused by the incomplete multi-modal data
. Specifically, apart from obtaining segmentation results from fused modal featu
res, we also segment each image modality individually from the corresponding enc
oded features. In this manner, each modal encoder is forced to learn discriminat
ive features, thus improving the representation ability of the fused features. R
emarkably, extensive experiments on BRATS2020, BRATS2018 and BRATS2015 datasets
demonstrate that our RFNet outperforms the state-of-the-art significantly.
****

Weakly Supervised Contrastive Learning
Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, Ch
ang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 10042-10051

Unsupervised visual representation learning has gained much attention from the c
omputer vision community because of the recent achievement of contrastive learni
ng. Most of the existing contrastive learning frameworks adopt the instance disc
rimination as the pretext task, which treating every single instance as a differ
ent class. However, such method will inevitably cause class collision problems,
which hurts the quality of the learned representation. Motivated by this observa
tion, we introduced a weakly supervised contrastive learning framework (WCL) to
tackle this issue. Specifically, our proposed framework is based on two projecti
on heads, one of which will perform the regular instance discrimination task. Th
e other head will use a graph-based method to explore similar samples and genera
te a weak label, then perform a supervised contrastive learning task based on th
e weak label to pull the similar images closer. We further introduced a K-Neares
t Neighbor based multi-crop strategy to expand the number of positive samples. E
xtensive experimental results demonstrate WCL improves the quality of self-super
vised representations across different datasets. Notably, we get a new state-of-
the-art result for semi-supervised learning. With only 1% and 10% labeled exampl
es, WCL achieves 65% and 72% ImageNet Top-1 Accuracy using ResNet50, which is ev
en higher than SimCLRv2 with ResNet101.
****

SLIM: Self-Supervised LiDAR Scene Flow and Motion Segmentation
Stefan Andreas Baur, David Josef Emmerichs, Frank Moosmann, Peter Pinggera, Björ
n Ommer, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on
 Computer Vision (ICCV), 2021, pp. 13126-13136
Recently, several frameworks for self-supervised learning of 3D scene flow on po
int clouds have emerged. Scene flow inherently separates every scene into multip
le moving agents and a large class of points following a single rigid sensor mot
ion. However, existing methods do not leverage this property of the data in thei
r self-supervised training routines which could improve and stabilize flow predi
ctions. Based on the discrepancy between a robust rigid ego-motion estimate and
a raw flow prediction, we generate a self-supervised motion segmentation signal.
 The predicted motion segmentation, in turn, is used by our algorithm to attend
to stationary points for aggregation of motion information in static parts of th
e scene. We learn our model end-to-end by backpropagating gradients through Kabs
ch's algorithm and demonstrate that this leads to accurate ego-motion which in t
urn improves the scene flow estimate. Using our method, we show state-of-the-art
 results across multiple scene flow metrics for different real-world datasets, s
howcasing the robustness and generalizability of this approach. We further analy
ze the performance gain when performing joint motion segmentation and scene flow
 in an ablation study. We also present a novel network architecture for 3D LiDAR
 scene flow which is capable of handling an order of magnitude more points durin
g training than previously possible.
****

Likelihood-Based Diverse Sampling for Trajectory Forecasting
Yecheng Jason Ma, Jeevana Priya Inala, Dinesh Jayaraman, Osbert Bastani; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, p
p. 13279-13288
Forecasting complex vehicle and pedestrian multi-modal distributions requires po
werful probabilistic approaches. Normalizing flows (NF) have recently emerged as
 an attractive tool to model such distributions. However, a key drawback is that
 independent samples drawn from a flow model often do not adequately capture all
 the modes in the underlying distribution. We propose Likelihood-Based Diverse S
ampling (LDS), a method for improving the quality and the diversity of trajector
y samples from a pre-trained flow model. Rather than producing individual sample
s, LDS produces a set of trajectories in one shot. Given a pre-trained forecasti
ng flow model, we train LDS using gradients from the model, to optimize an objec
tive function that rewards high likelihood for individual trajectories in the pr
edicted set, together with high spatial separation among trajectories. LDS outpe
rforms state-of-art post-hoc neural diverse forecasting methods for various pre-
trained flow models as well as conditional variational autoencoder (CVAE) models
. Crucially, it can also be used for transductive trajectory forecasting, where

the diverse forecasts are trained on-the-fly on unlabeled test examples. LDS is easy to implement, and we show that it offers a simple plug-in improvement over baselines on two challenging benchmarks. Code is at: https://github.com/JasonMa2016/LDS

****

In Defense of Scene Graphs for Image Captioning

Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, Truong Q. Nguyen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1407-1416

The mainstream image captioning models rely on Convolutional Neural Network (CNN) image features to generate captions via recurrent models. Recently, image scene graphs have been used to augment captioning models so as to leverage their structural semantics such as object entities, relationships and attributes. Several studies have noted that naive use of scene graphs from a black-box scene graph generator harms image captioning performance, and scene graph-based captioning models have to incur the overhead of explicit use of image features to generate decent captions. Addressing these challenges, we propose a framework, SG2Caps, that utilizes only the scene graph labels for competitive image captioning performance. The basic idea is to close the semantic gap between two scene graphs - one derived from the input image and the other one from its caption. In order to achieve this, we leverage the spatial location of objects and the Human-Object-Interaction (HOI) labels as an additional HOI graph. Our framework outperforms existing scene graph-only captioning models by a large margin indicating scene graphs as a promising representation for image captioning. Direct utilization of the scene graph labels avoids expensive graph convolutions over high-dimensional CNN features resulting in 49% fewer trainable parameters. The code is available at: https://github.com/Kien085/SG2Caps.

****

Rank & Sort Loss for Object Detection and Instance Segmentation

Kemal Oksuz, Baris Can Cam, Emre Akbas, Sinan Kalkan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3009-3018

We propose Rank & Sort (RS) Loss, a ranking-based loss function to train deep object detection and instance segmentation methods (i.e. visual detectors). RS Loss supervises the classifier, a sub-network of these methods, to rank each positive above all negatives as well as to sort positives among themselves with respect to (wrt.) their localisation qualities (e.g. Intersection-over-Union - IoU). To tackle the non-differentiable nature of ranking and sorting, we reformulate the incorporation of error-driven update with backpropagation as Identity Update, which enables us to model our novel sorting error among positives. With RS Loss, we significantly simplify training: (i) Thanks to our sorting objective, the positives are prioritized by the classifier without an additional auxiliary head (e.g. for centerness, IoU, mask-IoU), (ii) due to its ranking-based nature, RS Loss is robust to class imbalance, and thus, no sampling heuristic is required, and (iii) we address the multi-task nature of visual detectors using tuning-free task-balancing coefficients. Using RS Loss, we train seven diverse visual detectors only by tuning the learning rate, and show that it consistently outperforms baselines: e.g. our RS Loss improves (i) Faster R-CNN by 3 box AP and aLRP Loss (ranking-based baseline) by 2 box AP on COCO dataset, (ii) Mask R-CNN with repeat factor sampling (RFS) by 3.5 mask AP (7 AP for rare classes) on LVIS dataset; and also outperforms all counterparts. Code is available at: https://github.com/kemaloksuz/RankSortLoss.

****

RANK-NOSH: Efficient Predictor-Based Architecture Search via Non-Uniform Successive Halving

Ruochen Wang, Xiangning Chen, Minhao Cheng, Xiaocheng Tang, Cho-Jui Hsieh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10377-10386

Predictor-based algorithms have achieved remarkable performance in the Neural Architecture Search (NAS) tasks. However, these methods suffer from high computation costs, as training the performance predictor usually requires training and ev

aluating hundreds of architectures from scratch. Previous works along this line mainly focus on reducing the number of architectures required to fit the predictor. In this work, we tackle this challenge from a different perspective - improve search efficiency by cutting down the computation budget of architecture training. We propose NOn-uniform Successive Halving (NOSH), a hierarchical scheduling algorithm that terminates the training of underperforming architectures early to avoid wasting budget. To effectively leverage the non-uniform supervision signals produced by NOSH, we formulate predictor-based architecture search as learning to rank with pairwise comparisons. The resulting method - RANK-NOSH, reduces the search budget by 5x while achieving competitive or even better performance than previous state-of-the-art predictor-based methods on various spaces and datasets.

****

Dual Path Learning for Domain Adaptation of Semantic Segmentation

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, Wenqiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9082-9091

Domain adaptation for semantic segmentation enables to alleviate the need for large-scale pixel-wise annotations. Recently, self-supervised learning (SSL) with a combination of image-to-image translation shows great effectiveness in adaptive segmentation. The most common practice is to perform SSL along with image translation to well align a single domain (the source or target). However, in this single-domain paradigm, unavoidable visual inconsistency raised by image translation may affect subsequent learning. In this paper, based on the observation that domain adaptation frameworks performed in the source and target domain are almost complementary in terms of image translation and SSL, we propose a novel dual path learning (DPL) framework to alleviate visual inconsistency. Concretely, DPL contains two complementary and interactive single-domain adaptation pipelines aligned in source and target domain respectively. The inference of DPL is extremely simple, only one segmentation model in the target domain is employed. Novel technologies such as dual path image translation and dual path adaptive segmentation are proposed to make two paths promote each other in an interactive manner. Experiments on GTA5->Cityscapes and SYNTHIA->Cityscapes scenarios demonstrate the superiority of our DPL model over the state-of-the-art methods.

****

Synthesis of Compositional Animations From Textual Descriptions

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, Philipp Slusallek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1396-1406

How can we animate 3D-characters from a movie script or move robots by simply telling them what we would like them to do?" How unstructured and complex can we make a sentence and still generate plausible movements from it?" These are questions that need to be answered in the long-run, as the field is still in its infancy. Inspired by these problems, we present a new technique for generating compositional actions, which handles complex input sentences. Our output is a 3D pose sequence depicting the actions in the input sentence. We propose a hierarchical two-stream sequential model to explore a finer joint-level mapping between natural language sentences and 3D pose sequences corresponding to the given motion. We learn two manifold representations of the motion, one each for the upper body and the lower body movements. Our model can generate plausible pose sequences for short sentences describing single actions as well as long complex sentences describing multiple sequential and compositional actions. We evaluate our proposed model on the publicly available KIT Motion-Language Dataset containing 3D pose data with human-annotated sentences. Experimental results show that our model advances the state-of-the-art on text-based motion synthesis in objective evaluations by a margin of 50%. Qualitative evaluations based on a user study indicate that our synthesized motions are perceived to be the closest to the ground-truth motion captures for both short and compositional sentences.

****

Dual Bipartite Graph Learning: A General Approach for Domain Adaptive Object Det

ection
Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, Yizhou Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2703-2712
Domain Adaptive Object Detection (DAOD) relieves the reliance on large-scale annotated data by transferring the knowledge learned from a labeled source domain to a new unlabeled target domain. Recent DAOD approaches resort to local feature alignment in virtue of domain adversarial training in conjunction with the ad-hoc detection pipelines to achieve feature adaptation. However, these methods are limited to adapt the specific types of object detectors and do not explore the cross-domain topological relations. In this paper, we first formulate DAOD as an open-set domain adaptation problem in which foregrounds (pixel or region) can be seen as the "known class", while backgrounds (pixel or region) are referred to as the "unknown class". To this end, we present a new and general perspective for DAOD named Dual Bipartite Graph Learning (DBGL), which captures the cross-domain interactions on both pixel-level and semantic-level via increasing the distinction between foregrounds and backgrounds and modeling the cross-domain dependencies among different semantic categories. Experiments reveal that the proposed DBGL in conjunction with one-stage and two-stage detectors exceeds the state-of-the-art performance on standard DAOD benchmarks.
****

Parametric Contrastive Learning
Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 715-724
In this paper, we propose Parametric Contrastive Learning (PaCo) to tackle long-tailed recognition. Based on theoretical analysis, we observe supervised contrastive loss tends to bias on high-frequency classes and thus increases the difficulty of imbalanced learning. We introduce a set of parametric class-wise learnable centers to rebalance from an optimization perspective. Further, we analyze our PaCo loss under a balanced setting. Our analysis demonstrates that PaCo can adaptively enhance the intensity of pushing samples of the same class close as more samples are pulled together with their corresponding centers and benefit hard example learning. Experiments on long-tailed CIFAR, ImageNet, Places, and iNaturalist 2018 manifest the new state-of-the-art for long-tailed recognition. On full ImageNet, models trained with PaCo loss surpass supervised contrastive learning across various ResNet backbones, e.g., our ResNet-200 achieves 81.8% top-1 accuracy. Our code is available at https://github.com/dvlab-research/Parametric-Contrastive-Learning.
****

A Simple Feature Augmentation for Domain Generalization
Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, Timothy M. Hospedales; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8886-8895
The topical domain generalization (DG) problem asks trained models to perform well on an unseen target domain with different data statistics from the source training domains. In computer vision, data augmentation has proven one of the most effective ways of better exploiting the source data to improve domain generalization. However, existing approaches primarily rely on image-space data augmentation, which requires careful augmentation design, and provides limited diversity of augmented data. We argue that feature augmentation is a more promising direction for DG. We find that an extremely simple technique of perturbing the feature embedding with Gaussian noise during training leads to a classifier with domain-generalization performance comparable to existing state of the art. To model more meaningful statistics reflective of cross-domain variability, we further estimate the full class-conditional feature covariance matrix iteratively during training. Subsequent joint stochastic feature augmentation provides an effective domain randomization method, perturbing features in the directions of intra-class/cross-domain variability. We verify our proposed method on three standard domain generalization benchmarks, Digit-DG, VLCS and PACS, and show it is outperforming or comparable to the state of the art in all setups, together with experimental

analysis to illustrate how our method works towards training a robust generalisable model.
****

Visual Transformers: Where Do Transformers Really Belong in Vision Models?
Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph E. Gonzalez, Kurt Keutzer, Peter Vajda; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 599-609
A recent trend in computer vision is to replace convolutions with transformers. However, the performance gain of transformers is attained at a steep cost, requiring GPU years and hundreds of millions of samples for training. This excessive resource usage compensates for a misuse of transformers: Transformers densely model relationships between its inputs -- ideal for late stages of a neural network, when concepts are sparse and spatially-distant, but extremely inefficient for early stages of a network, when patterns are redundant and localized. To address these issues, we leverage the respective strengths of both operations, building convolution-transformer hybrids. Critically, in sharp contrast to pixel-space transformers, our Visual Transformer (VT) operates in a semantic token space, judiciously attending to different image parts based on context. Our VTs significantly outperforms baselines: On ImageNet, our VT-ResNets outperform convolution-only ResNet by 4.6 to 7 points and transformer-only ViT-B by 2.6 points with 2.5 times fewer FLOPs, 2.1 times fewer parameters. For semantic segmentation on LIP and COCO-stuff, VT-based feature pyramid networks (FPN) achieve 0.35 points higher mIoU while reducing the FPN module's FLOPs by 6.5x.
****

Is Pseudo-Lidar Needed for Monocular 3D Object Detection?
Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, Adrien Gaidon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3142-3152
Recent progress in 3D object detection from single images leverages monocular depth estimation as a way to produce 3D pointclouds, turning cameras into pseudo-lidar sensors. These two-stage detectors improve with the accuracy of the intermediate depth estimation network, which can itself be improved without manual labels via large-scale self-supervised learning. However, they tend to suffer from overfitting more than end-to-end methods, are more complex, and the gap with similar lidar-based detectors remains significant. In this work, we propose an end-to-end, single stage, monocular 3D object detector, DD3D, that can benefit from depth pre-training like pseudo-lidar methods, but without their limitations. Our architecture is designed for effective information transfer between depth estimation and 3D detection, allowing us to scale with the amount of unlabeled pre-training data. Our method achieves state-of-theart results on two challenging benchmarks, with 16.34% and 9.28% AP for Cars and Pedestrians (respectively) on the KITTI-3D benchmark, and 41.5% mAP on NuScenes.
****

TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization
Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, Qixiang Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2886-2895
Weakly supervised object localization (WSOL) is a challenging problem when given image category labels but requires to learn object localization models. Optimizing a convolutional neural network (CNN) for classification tends to activate local discriminative regions while ignoring complete object extent, causing the partial activation issue. In this paper, we argue that partial activation is caused by the intrinsic characteristics of CNN, where the convolution operations produce local receptive fields and experience difficulty to capture long-range feature dependency among pixels. We introduce the token semantic coupled attention map (TS-CAM) to take full advantage of the self-attention mechanism in visual transformer for long-range dependency extraction. TS-CAM first splits an image into a sequence of patch tokens for spatial embedding, which produce attention maps o

f long-range visual dependency to avoid partial activation. TS-CAM then re-alloc
ates category-related semantics for patch tokens, enabling each of them to be aw
are of object categories. TS-CAM finally couples the patch tokens with the seman
tic-agnostic attention map to achieve semantic-aware localization. Experiments o
n the ILSVRC/CUB-200-2011 datasets show that TS-CAM outperforms its CNN-CAM coun
terparts by7.1%/27.1%for WSOL, achieving state-of-the-art performance. Code is a
vailable at https://github.com/vasgaowei/TS-CAM
****

Geometry-Based Distance Decomposition for Monocular 3D Object Detection
Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, Tae-Kyun Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15172-15181
Monocular 3D object detection is of great significance for autonomous driving bu
t remains challenging. The core challenge is to predict the distance of objects
in the absence of explicit depth information. Unlike regressing the distance as
a single variable in most existing methods, we propose a novel geometry-based di
stance decomposition to recover the distance by its factors. The decomposition f
actors the distance of objects into the most representative and stable variables
, i.e. the physical height and the projected visual height in the image plane. M
oreover, the decomposition maintains the self-consistency between the two height
s, leading to robust distance prediction when both predicted heights are inaccur
ate. The decomposition also enables us to trace the causes of the distance uncer
tainty for different scenarios. Such decomposition makes the distance prediction
 interpretable, accurate, and robust. Our method directly predicts 3D bounding b
oxes from RGB images with a compact architecture, making the training and infere
nce simple and efficient. The experimental results show that our method achieves
 the state-of-the-art performance on the monocular 3D Object Detection and Bird'
s Eye View tasks of the KITTI dataset, and can generalize to images with differe
nt camera intrinsics.
****

Deep 3D Mask Volume for View Synthesis of Dynamic Scenes
Kai-En Lin, Lei Xiao, Feng Liu, Guowei Yang, Ravi Ramamoorthi; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1749-17
58
Image view synthesis has seen great success in reconstructing photorealistic vis
uals, thanks to deep learning and various novel representations. The next key st
ep in immersive virtual experiences is view synthesis of dynamic scenes. However
, several challenges exist due to the lack of high-quality training datasets, an
d the additional time dimension for videos of dynamic scenes. To address this is
sue, we introduce a multi-view video dataset, captured with a custom 10-camera r
ig in 120FPS. The dataset contains 96 high-quality scenes showing various visual
 effects and human interactions in outdoor scenes. We develop a new algorithm, D
eep 3D Mask Volume, which enables temporally-stable view extrapolation from bino
cular videos of dynamic scenes, captured by static cameras. Our algorithm addres
ses the temporal inconsistency of disocclusions by identifying the error-prone a
reas with a 3D mask volume, and replaces them with static background observed th
roughout the video. Our method enables manipulation in 3D space as opposed to si
mple 2D masks. We demonstrate better temporal stability than frame-by-frame stat
ic view synthesis methods, or those that use 2D masks. The resulting view synthe
sis videos show minimal flickering artifacts and allow for larger translational
movements.
****

Unified Questioner Transformer for Descriptive Question Generation in Goal-Orien
ted Visual Dialogue
Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, Mi
chita Imai; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2021, pp. 1898-1907
Building an interactive artificial intelligence that can ask questions about the
 real world is one of the biggest challenges for vision and language problems. I
n particular, goal-oriented visual dialogue, where the aim of the agent is to se

ek information by asking questions during a turn-taking dialogue, has been gaining scholarly attention recently. While several existing models based on the GuessWhat?! dataset have been proposed, the Questioner typically asks simple category-based questions or absolute spatial questions. This might be problematic for complex scenes where the objects share attributes or in cases where descriptive questions are required to distinguish objects. In this paper, we propose a novel Questioner architecture, called Unified Questioner Transformer (UniQer), for descriptive question generation with referring expressions. In addition, we build a goal-oriented visual dialogue task called CLEVR Ask. It synthesizes complex scenes that require the Questioner to generate descriptive questions. We train our model with two variants of CLEVR Ask datasets. The results of the quantitative and qualitative evaluations show that UniQer outperforms the baseline.

****

Human Trajectory Prediction via Counterfactual Analysis

Guangyi Chen, Junlong Li, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9824-9833

Forecasting human trajectories in complex dynamic environments plays a critical role in autonomous vehicles and intelligent robots. Most existing methods learn to predict future trajectories by behavior clues from history trajectories and interaction clues from environments. However, the inherent bias between training and deployment environments is ignored. Hence, we propose a counterfactual analysis method for human trajectory prediction to investigate the causality between the predicted trajectories and input clues and alleviate the negative effects brought by environment bias. We first build a causal graph for trajectory forecasting with history trajectory, future trajectory, and the environment interactions. Then, we cut off the inference from the environment to trajectory by constructing the counterfactual intervention on the trajectory itself. Finally, we compare the factual and counterfactual trajectory clues to alleviate the effects of environment bias and highlight the trajectory clues. Our counterfactual analysis is a plug-and-play module that can be applied to any baseline prediction methods including RNN- and CNN-based ones. We show that our method achieves consistent improvement for different baselines and obtains state-of-the-art results on public pedestrian trajectory forecasting benchmarks.

****

Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-Identification

Yongming Rao, Guangyi Chen, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1025-1034

Attention mechanism has demonstrated great potential in fine-grained visual recognition tasks. In this paper, we present a counterfactual attention learning method to learn more effective attention based on causal inference. Unlike most existing methods that learn visual attention based on conventional likelihood, we propose to learn the attention with counterfactual causality, which provides a tool to measure the attention quality and a powerful supervisory signal to guide the learning process. Specifically, we analyze the effect of the learned visual attention on network prediction through counterfactual intervention and maximize the effect to encourage the network to learn more useful attention for fine-grained image recognition. Empirically, we evaluate our method on a wide range of fine-grained visual recognition tasks where attention plays a crucial role, including fine-grained image categorization, person re-identification, and vehicle re-identification. The consistent improvement on all benchmarks demonstrates the effectiveness of our method.

****

Effectively Leveraging Attributes for Visual Similarity

Samarth Mishra, Zhongping Zhang, Yuan Shen, Ranjitha Kumar, Venkatesh Saligrama, Bryan A. Plummer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1015-1024

Measuring similarity between two images often requires performing complex reasoning along different axes (e.g., color, texture, or shape). Insights into what might be important for measuring similarity can can be provided by annotated attri

butes, but prior work tends to view these annotations as complete, resulting in them using a simplistic approach of predicting attributes on single images, which are, in turn, used to measure similarity. However, it is impractical for a dataset to fully annotate every attribute that may be important. Thus, only representing images based on these incomplete annotations may miss out on key information. To address this issue, we propose the Pairwise Attribute-informed similarity Network (PAN), which breaks similarity learning into capturing similarity conditions and relevance scores from a joint representation of two images. This enables our model to identify that two images contain the same attribute, but can have it deemed irrelevant (e.g., due to fine-grained differences between them) and ignored for measuring similarity between the two images. Notably, while prior methods of using attribute annotations are often unable to outperform prior art, PAN obtains a 4-9% improvement on compatibility prediction between clothing items on Polyvore Outfits, a 5% gain on few shot classification of images using Caltech-UCSD Birds (CUB), and over 1% boost to Recall@1 on In-Shop Clothes Retrieval.
****

Anticipative Video Transformer
Rohit Girdhar, Kristen Grauman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13505-13515
We propose Anticipative Video Transformer (AVT), an end-to-end attention-based video modeling architecture that attends to the previously observed video in order to anticipate future actions. We train the model jointly to predict the next action in a video sequence, while also learning frame feature encoders that are predictive of successive future frames' features. Compared to existing temporal aggregation strategies, AVT has the advantage of both maintaining the sequential progression of observed actions while still capturing long-range dependencies--both critical for the anticipation task. Through extensive experiments, we show that AVT obtains the best reported performance on four popular action anticipation benchmarks: EpicKitchens-55, EpicKitchens-100, EGTEA Gaze+, and 50-Salads; and it wins first place in the EpicKitchens-100 CVPR'21 challenge.
****

Semantically Robust Unpaired Image Translation for Data With Unmatched Semantics Statistics
Zhiwei Jia, Bodi Yuan, Kangkang Wang, Hong Wu, David Clifford, Zhiqiang Yuan, Hao Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14273-14283
Many applications of unpaired image-to-image translation require the input contents to be preserved semantically during translations. Unaware of the inherently unmatched semantics distributions between source and target domains, existing distribution matching methods (i.e., GAN-based) can give undesired solutions. In specific, although producing visually reasonable outputs, the learned models usually flip the semantics of the inputs. To tackle this without using extra supervisions, we propose to enforce the translated outputs to be semantically invariant w.r.t. small perceptual variations of the inputs, a property we call ""semantic robustness"". By optimizing a robustness loss w.r.t. multi-scale feature space perturbations of the inputs, our method effectively reduces semantics flipping and produces translations that outperform existing methods both quantitatively and qualitatively.
****

Progressive Seed Generation Auto-Encoder for Unsupervised Point Cloud Learning
Juyoung Yang, Pyunghwan Ahn, Doyeon Kim, Haeil Lee, Junmo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6413-6422
With the development of 3D scanning technologies, 3D vision tasks have become a popular research area. Owing to the large amount of data acquired by sensors, unsupervised learning is essential for understanding and utilizing point clouds without an expensive annotation process. In this paper, we propose a novel framework and an effective auto-encoder architecture named "PSG-Net" for reconstruction-based learning of point clouds. Unlike existing studies that used fixed or random 2D points, our framework generates input-dependent point-wise features for th

e latent point set. PSG-Net uses the encoded input to produce point-wise feature
s through the seed generation module and extracts richer features in multiple st
ages with gradually increasing resolution by applying the seed feature propagati
on module progressively. We prove the effectiveness of PSG-Net experimentally; P
SG-Net shows state-of-the-art performances in point cloud reconstruction and uns
upervised classification, and achieves comparable performance to counterpart met
hods in supervised completion.
****

Waypoint Models for Instruction-Guided Navigation in Continuous Environments
Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, Oleksandr Maksymets; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
, pp. 15162-15171
Little inquiry has explicitly addressed the role of action spaces in language-gu
ided visual navigation -- either in terms of its effect on navigation success or
 the efficiency with which a robotic agent could execute the resulting trajector
y. Building on the recently released VLN-CE setting for instruction following in
 continuous environments, we develop a class of language-conditioned waypoint pr
ediction networks to examine this question. We vary the expressivity of these mo
dels to explore a spectrum between low-level actions and continuous waypoint pre
diction. We measure task performance and estimated execution time on a profiled
LoCoBot robot. We find more expressive models result in simpler, faster to execu
te trajectories, but lower-level actions can achieve better navigation metrics b
y approximating shortest paths better. Further, our models outperform prior work
 in VLN-CE and set a new state-of-the-art on the public leaderboard -- increasin
g success rate by 4% with our best model on this challenging task.
****

Rethinking Preventing Class-Collapsing in Metric Learning With Margin-Based Loss
es
Elad Levi, Tete Xiao, Xiaolong Wang, Trevor Darrell; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 10316-10325
Metric learning seeks perceptual embeddings where visually similar instances are
 close and dissimilar instances are apart, but learned representations can be su
b-optimal when the distribution of intra-class samples is diverse and distinct s
ub-clusters are present. Although theoretically with optimal assumptions, margin
-based losses such as the triplet loss and margin loss have a diverse family of
solutions. We theoretically prove and empirically show that under reasonable noi
se assumptions, margin-based losses tend to project all samples of a class with
various modes onto a single point in the embedding space, resulting in a class c
ollapse that usually renders the space ill-sorted for classification or retrieva
l. To address this problem, we propose a simple modification to the embedding lo
sses such that each sample selects its nearest same-class counterpart in a batch
 as the positive element in the tuple. This allows for the presence of multiple
sub-clusters within each class. The adaptation can be integrated into a wide ran
ge of metric learning losses. The proposed sampling method demonstrates clear be
nefits on various fine-grained image retrieval datasets over a variety of existi
ng losses; qualitative retrieval results show that samples with similar visual p
atterns are indeed closer in the embedding space.
****

HiNet: Deep Image Hiding by Invertible Network
Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, Zhenyu Guan; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4733-4742
Image hiding aims to hide a secret image into a cover image in an imperceptible
way, and then recover the secret image perfectly at the receiver end. Capacity,
invisibility and security are three primary challenges in image hiding task. Thi
s paper proposes a novel invertible neural network (INN) based framework, HiNet,
 to simultaneously overcome the three challenges in image hiding. For large capa
city, we propose an inverse learning mechanism by simultaneously learning the im
age concealing and revealing processes. Our method is able to achieve the concea
ling of a full-size secret image into a cover image with the same size. For high
 invisibility, instead of pixel domain hiding, we propose to hide the secret inf

ormation in wavelet domain. Furthermore, we propose a new low-frequency wavelet loss to constrain that secret information is hidden in high-frequency wavelet sub-bands, which significantly improves the hiding security. Experimental results show that our HiNet significantly outperforms other state-of-the-art image hiding methods, with more than 10 dB PSNR improvement in secret image recovery on ImageNet, COCO and DIV2K datasets. Codes are available at https://github.com/TomTomTommi/HiNet.

****

Rotation Averaging in a Split Second: A Primal-Dual Method and a Closed-Form for Cycle Graphs

Gabriel Moreira, Manuel Marques, João Paulo Costeira; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5452-5460

A cornerstone of geometric reconstruction, rotation averaging seeks the set of absolute rotations that optimally explains a set of measured relative orientations between them. In spite of being an integral part of bundle adjustment and structure-from-motion, averaging rotations is both a nonconvex and high-dimensional optimization problem. In this paper, we address it from a maximum likelihood estimation standpoint and make a twofold contribution. Firstly, we set forth a novel initialization-free primal-dual method which we show empirically to converge to the global optimum. Further, we derive what is to our knowledge, the first optimal closed-form solution for rotation averaging in cycle graphs and contextualize this result within spectral graph theory. Our proposed methods achieve a significant gain both in precision and performance.

****

End-to-End Robust Joint Unsupervised Image Alignment and Clustering

Xiangrui Zeng, Gregory Howe, Min Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3854-3866

Computing dense pixel-to-pixel image correspondences is a fundamental task of computer vision. Often, the objective is to align image pairs from the same semantic category for manipulation or segmentation purposes. Despite achieving superior performance, existing deep learning alignment methods cannot cluster images; consequently, clustering and pairing images needed to be a separate laborious and expensive step. Given a dataset with diverse semantic categories, we propose a multi-task model, Jim-Net, that can directly learn to cluster and align images without any pixel-level or image-level annotations. We design a pair-matching alignment unsupervised training algorithm that selectively matches and aligns image pairs from the clustering branch. Our unsupervised Jim-Net achieves comparable accuracy with state-of-the-art supervised methods on benchmark 2D image alignment dataset PF-PASCAL. Specifically, we apply Jim-Net to cryo-electron tomography, a revolutionary 3D microscopy imaging technique of native subcellular structures. After extensive evaluation on seven datasets, we demonstrate that Jim-Net enables systematic discovery and recovery of representative macromolecular structures in situ, which is essential for revealing molecular mechanisms underlying cellular functions. To our knowledge, Jim-Net is the first end-to-end model that can simultaneously align and cluster images, which significantly improves the performance as compared to performing each task alone.

****

BabelCalib: A Universal Approach to Calibrating Central Cameras

Yaroslava Lochman, Kostiantyn Liepieshov, Jianhui Chen, Michal Perdoch, Christopher Zach, James Pritts; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15253-15262

Existing calibration methods occasionally fail for large field-of-view cameras due to the non-linearity of the underlying problem and the lack of good initial values for all parameters of the used camera model. This might occur because a simpler projection model is assumed in an initial step, or a poor initial guess for the internal parameters is pre-defined. A lot of the difficulties of general camera calibration lie in the use of a forward projection model. We side-step these challenges by first proposing a solver to calibrate the parameters in terms of a back-projection model and then regress the parameters for a target forward model. These steps are incorporated in a robust estimation framework to cope with

outlying detections. Extensive experiments demonstrate that our approach is very reliable and returns the most accurate calibration parameters as measured on the downstream task of absolute pose estimation on test sets. The code is released at https://github.com/ylochman/babelcalib

****

Curious Representation Learning for Embodied Intelligence

Yilun Du, Chuang Gan, Phillip Isola; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10408-10417

Self-supervised visual representation learning has achieved remarkable success in recent years. By subverting the need for supervised labels, such approaches are able to utilize the numerous unlabeled images that exist on the Internet and in photographic datasets. Yet to build truly intelligent agents, we must construct representation learning algorithms that can learn not only from datasets but also learn in environments. An agent in a natural environment will not typically be fed curated data. Instead, it must explore its environment to acquire the data it will learn from. We propose a framework, curious representation learning (CRL), which jointly learns a reinforcement learning policy and a visual representation model. The policy is trained to maximize the error of the representation learner, and in doing so is incentivized to explore its environment. At the same time, the learned representation becomes stronger and stronger as the policy feeds it ever harder data to learn from. Our learned embodied representations enable promising transfer to downstream embodied semantic and language-guided navigation, performing better or comparable to ImageNet pretraining without using any supervision at all. In addition, despite being trained in simulation, our learned representations can obtain interpretable results on real images.

****

Multi-Modal Multi-Action Video Recognition

Zhensheng Shi, Ju Liang, Qianqian Li, Haiyong Zheng, Zhaorui Gu, Junyu Dong, Bing Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13678-13687

Multi-action video recognition is much more challenging due to the requirement to recognize multiple actions co-occurring simultaneously or sequentially. Modeling multi-action relations is beneficial and crucial to understand videos with multiple actions, and actions in a video are usually presented in multiple modalities. In this paper, we propose a novel multi-action relation model for videos, by leveraging both relational graph convolutional networks (GCNs) and video multi-modality. We first build multi-modal GCNs to explore modality-aware multi-action relations, fed by modality-specific action representation as node features, e.g., spatiotemporal features learned by 3D convolutional neural network (CNN), audio and textual embeddings queried from respective feature lexicons. We then joint both multi-modal CNN-GCN models and multi-modal feature representations for learning better relational action predictions. Ablation study, multi-action relation visualization, and boosts analysis, all show efficacy of our multi-modal multi-action relation modeling. Also our method achieves state-of-the-art performance on large-scale multi-action M-MiT benchmark. Our code is made publicly available at https://github.com/zhenglab/multi-action-video.

****

Cross-Patch Graph Convolutional Network for Image Denoising

Yao Li, Xueyang Fu, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4651-4660

Recently, deep learning-based image denoising methods have achieved significant improvements over traditional methods. Due to the hardware limitation, most deep learning-based image denoising methods utilize cropped small patches to train a convolutional neural network to infer the clean images. However, the real noisy images in practical are mostly of high resolution rather than the cropped small patches and the vanilla training strategies ignore the cross-patch contextual dependency in the whole image. In this paper, we propose Cross-Patch Net (CPNet), which is the first deep- learning-based real image denoising method for HR (high resolution) input. Furthermore, we design a novel loss guided by the noise level map to obtain better performance. Compared with the vanilla patch-based train

ing strategies, our approach effectively exploits the cross-patch contextual dep
endency. effective method to generate realistic sRGB noisy images from their cor
responding clean sRGB images for denoiser training. Denoising experiments on rea
l-world sRGB images show the effectiveness of the proposed method. More importan
tly, our method achieves state-of-the-art performance on practical sRGB noisy im
age denoising.
****

ISNet: Integrate Image-Level and Semantic-Level Context for Semantic Segmentatio
n
Zhenchao Jin, Bin Liu, Qi Chu, Nenghai Yu; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2021, pp. 7189-7198
Co-occurrent visual pattern makes aggregating contextual information a common pa
radigm to enhance the pixel representation for semantic image segmentation. The
existing approaches focus on modeling the context from the perspective of the wh
ole image, i.e., aggregating the image-level contextual information. Despite imp
ressive, these methods weaken the significance of the pixel representations of t
he same category, i.e., the semantic-level contextual information. To address th
is, this paper proposes to augment the pixel representations by aggregating the
image-level and semantic-level contextual information, respectively. First, an i
mage-level context module is designed to capture the contextual information for
each pixel in the whole image. Second, we aggregate the representations of the s
ame category for each pixel where the category regions are learned under the sup
ervision of the ground-truth segmentation. Third, we compute the similarities be
tween each pixel representation and the image-level contextual information, the
semantic-level contextual information, respectively. At last, a pixel representa
tion is augmented by weighted aggregating both the image-level contextual inform
ation and the semantic-level contextual information with the similarities as the
 weights. Integrating the image-level and semantic-level context allows this pap
er to report state-of-the-art accuracy on four benchmarks, i.e., ADE20K, LIP, CO
COStuff and Cityscapes.
****

Body-Face Joint Detection via Embedding and Head Hook
Junfeng Wan, Jiangfan Deng, Xiaosong Qiu, Feng Zhou; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2021, pp. 2959-2968
Detecting pedestrians and their associated faces jointly is a challenging task.O
n one hand, body or face could be absent because of occlusion or non-frontal hum
an pose.On the other hand, the association becomes difficult or even miss-leadin
g in crowded scenes due to the lack of strong correlational evidence. This paper
 proposes Body-Face Joint (BFJ) detector, a novel framework for detecting bodies
 and their faces with accurate correspondance. We follow the classical multi-cla
ss detector design by detecting body and face in parallel but with two key contr
ibutions. First, we propose an Embedding Matching Loss (EML) to learn an associa
tive embedding for matching body and face of the same person. Second, we introdu
ce a novel concept, "head hook", to bridge the gap of matching body and faces sp
atially. With the new semantical and geometrical sources of information, BFJ gre
atly reduces the difficulty of detecting body and face in pairs. Since the probl
em is unexplored yet, we design a new metric named log-average miss matching rat
e (mMR^ -2) to evaluate the association performance and extend the CrowdHuman a
nd CityPersons benchmarks by annotating each face box. Experiments show that our
 BFJ detector can maintain state-of-the-art performance in pedestrian detection
on both one-stage and two-stage structures while greatly outperform various body
-face association strategies. Code and datasets will be released soon.
****

Enhancing Self-Supervised Video Representation Learning via Multi-Level Feature
Optimization
Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, Weiy
ao Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2021, pp. 7990-8001
The crux of self-supervised video representation learning is to build general fe
atures from unlabeled videos. However, most recent works have mainly focused on

high-level semantics and neglected lower-level representations and their tempora
l relationship which are crucial for general video understanding. To address the
se challenges, this paper proposes a multi-level feature optimization framework
to improve the generalization and temporal modeling ability of learned video rep
resentations. Concretely, high-level features obtained from naive and prototypic
al contrastive learning are utilized to build distribution graphs, guiding the p
rocess of low-level and mid-level feature learning. We also devise a simple temp
oral modeling module from multi-level features to enhance motion pattern learnin
g. Experiments demonstrate that multi-level feature optimization with the graph
constraint and temporal modeling can greatly improve the representation ability
in video understanding. Code is available at https://github.com/shvdiwnkozbw/Vid
eo-Representation-via-Multi-level-Optimization.
****

LIGA-Stereo: Learning LiDAR Geometry Aware Representations for Stereo-Based 3D D
etector
Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, Hongsheng Li; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3153-3163
Stereo-based 3D detection aims at detecting 3D object bounding boxes from stereo
 images using intermediate depth maps or implicit 3D geometry representations, w
hich provides a low-cost solution for 3D perception. However, its performance is
 still inferior compared with LiDAR-based detection algorithms. To detect and lo
calize accurate 3D bounding boxes, LiDAR-based models can encode accurate object
 boundaries and surface normal directions from LiDAR point clouds. However, the
detection results of stereo-based detectors are easily affected by the erroneous
 depth features due to the limitation of stereo matching. To solve the problem,
we propose LIGA-Stereo (LiDAR Geometry Aware Stereo Detector) to learn stereo-ba
sed 3D detectors under the guidance of high-level geometry-aware representations
 of LiDAR-based detection models. In addition, we found existing voxel-based ste
reo detectors failed to learn semantic features effectively from indirect 3D sup
ervisions. We attach an auxiliary 2D detection head to provide direct 2D semanti
c supervisions. Experiment results show that the above two strategies improved t
he geometric and semantic representation capabilities. Compared with the state-o
f-the-art stereo detector, our method has improved the 3D detection performance
of cars, pedestrians, cyclists by 10.44%, 5.69%, 5.97% mAP respectively on the o
fficial KITTI benchmark. The gap between stereo-based and LiDAR-based 3D detecto
rs is further narrowed.
****

Semi-Supervised Semantic Segmentation With Pixel-Level Contrastive Learning From
 a Class-Wise Memory Bank
Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, Ana C. Murillo; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
1, pp. 8219-8228
This work presents a novel approach for semi-supervised semantic segmentation. T
he key element of this approach is our contrastive learning module that enforces
 the segmentation network to yield similar pixel-level feature representations f
or same-class samples across the whole dataset. To achieve this, we maintain a m
emory bank which is continuously updated with relevant and high-quality feature
vectors from labeled data. In an end-to-end training, the features from both lab
eled and unlabeled data are optimized to be similar to same-class samples from t
he memory bank. Our approach not only outperforms the current state-of-the-art f
or semi-supervised semantic segmentation but also for semi-supervised domain ada
ptation on well-known public benchmarks, with larger improvements on the most ch
allenging scenarios, i.e., less available labeled data. Code is available at htt
ps://github.com/Shathe/SemiSeg-Contrastive
****

End-to-End Urban Driving by Imitating a Reinforcement Learning Coach
Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, Luc Van Gool; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp.
 15222-15232
End-to-end approaches to autonomous driving commonly rely on expert demonstratio

ns. Although humans are good drivers, they are not good coaches for end-to-end a lgorithms that demand dense on-policy supervision. On the contrary, automated ex perts that leverage privileged information can efficiently generate large scale on-policy and off-policy demonstrations. However, existing automated experts for urban driving make heavy use of hand-crafted rules and perform suboptimally eve n on driving simulators, where ground-truth information is available. To address these issues, we train a reinforcement learning expert that maps bird's-eye vie w images to continuous low-level actions. While setting a new performance upper- bound on CARLA, our expert is also a better coach that provides informative supe rvision signals for imitation learning agents to learn from. Supervised by our r einforcement learning coach, a baseline end-to-end agent with monocular camera-i nput achieves expert-level performance. Our end-to-end agent achieves a 78% succ ess rate while generalizing to a new town and new weather on the NoCrash-dense b enchmark and state-of-the-art performance on the more challenging CARLA LeaderBo ard.

****

Interpolation-Aware Padding for 3D Sparse Convolutional Neural Networks

Yu-Qi Yang, Peng-Shuai Wang, Yang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7467-7475

Sparse voxel-based 3D convolutional neural networks (CNNs) are widely used for v arious 3D vision tasks. Sparse voxel-based 3D CNNs create sparse non-empty voxel s from input point clouds and perform standard convolution operations on them on ly. We propose a simple and effective padding scheme --- interpolation-aware pad ding to pad a few empty voxels adjacent to the non-empty voxels and involving th em in the CNN computation so that all neighboring voxels exist when computing po int-wise features via the trilinear interpolation. For fine-grained 3D vision ta sks where point-wise features are essential, like semantic segmentation and 3D d etection, our network achieves higher prediction accuracy than the existing netw orks using the nearest neighbor interpolation or normalized trilinear interpolat ion with the zero-padding or the octree-padding scheme. Through extensive compar isons on various 3D segmentation and detection tasks, we demonstrate the superio rity of 3D sparse CNNs with our sparse padding scheme in conjunction with featur e interpolation.

****

Active Learning for Lane Detection: A Knowledge Distillation Approach

Fengchao Peng, Chao Wang, Jianzhuang Liu, Zhen Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15152-15161

Lane detection is a key task for autonomous driving vehicles. Currently, lane de tection relies on a huge amount of annotated images, which is a heavy burden. Ac tive learning has been proposed to reduce annotation in many computer vision tas ks, but no effort has been made for lane detection. Through experiments, we find that existing active learning methods perform poorly for lane detection, and th e reasons are twofold. On one hand, most methods evaluate data uncertainties bas ed on entropy, which is undesirable in lane detection because it encourages to s elect images with very few lanes or even no lane at all. On the other hand, exis ting methods are not aware of the noise of lane annotations, which is caused by heavy occlusion and unclear lane marks. In this paper, we build a novel knowledg e distillation framework and evaluate the uncertainty of images based on the kno wledge learnt by the student model. We show that the proposed uncertainty metric overcomes the above two problems. To reduce data redundancy, we explore the inf luence sets of image samples, and propose a new diversity metric for data select ion. Finally we incorporate the uncertainty and diversity metrics, and develop a greedy algorithm for data selection. The experiments show that our method achie ves new state-of-the-art on the lane detection benchmarks. In addition, we exten d this method to common 2D object detection and the results show that it is also effective.

****

Once Quantization-Aware Training: High Performance Extremely Low-Bit Architectur e Search

Mingzhu Shen, Feng Liang, Ruihao Gong, Yuhang Li, Chuming Li, Chen Lin, Fengwei

Yu, Junjie Yan, Wanli Ouyang; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2021, pp. 5340-5349
Quantization Neural Networks (QNN) have attracted a lot of attention due to thei
r high efficiency. To enhance the quantization accuracy, prior works mainly focu
s on designing advanced quantization algorithms but still fail to achieve satisf
actory results under the extremely low-bit case. In this work, we take an archit
ecture perspective to investigate the potential of high-performance QNN. Therefo
re, we propose to combine Network Architecture Search methods with quantization
to enjoy the merits of the two sides. However, a naive combination inevitably fa
ces unacceptable time consumption or unstable training problem. To alleviate the
se problems, we first propose the joint training of architecture and quantizatio
n with a shared step size to acquire a large number of quantized models. Then a
bit-inheritance scheme is introduced to transfer the quantized models to the low
er bit, which further reduces the time cost and meanwhile improves the quantizat
ion accuracy. Equipped with this overall framework, dubbed as Once Quantization-
Aware Training (OQAT), our searched model family, OQATNets, achieves a new state
-of-the-art compared with various architectures under different bit-widths. In p
articular, OQAT-2bit-M achieves 61.6% ImageNet Top-1 accuracy, outperforming 2-b
it counterpart MobileNetV3 by a large margin of 9% with 10% less computation cos
t. A series of quantization-friendly architectures are identified easily and ext
ensive analysis can be made to summarize the interaction between quantization an
d neural architectures. Codes and models are released at https://github.com/LaVi
eEnRoseSMZ/OQA
****

Learn To Match: Automatic Matching Network Design for Visual Tracking
Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, Weiming Hu; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13339-13348
Siamese tracking has achieved groundbreaking performance in recent years, where
the essence is the efficient matching operator cross-correlation and its variant
s. Besides the remarkable success, it is important to note that the heuristic ma
tching network design relies heavily on expert experience. Moreover, we experime
ntally find that one sole matching operator is difficult to guarantee stable tra
cking in all challenging environments. Thus, in this work, we introduce six nove
l matching operators, namely Concatenation, Pointwise-Addition, Pairwise-Relatio
n, FiLM, Simple-Transformer and Transductive-Guidance, to explore more feasibili
ty on matching operator selection. The analyses reveal these operators' selectiv
e adaptability on different environment degradation types, which inspires us to
combine them to explore complementary features. To this end, we propose binary c
hannel manipulation (BCM) to search for the optimal combination of these operato
rs. BCM determines to retrain or discard one operator by learning its contributi
on to other tracking steps. By inserting the learned matching networks to a stro
ng baseline tracker Ocean, our model achieves favorable gains by 67.2 -> 71.4, 5
2.6 -> 58.3, 70.3 -> 76.0 AUC on OTB100, LaSOT, and TrackingNet, respectively. N
otably, Our tracker runs at real-time speed of 50 / 100 FPS using PyTorch / Tens
orRT.
****

Oriented R-CNN for Object Detection
Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, Junwei Han; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3520-352
9
Current state-of-the-art two-stage detectors generate oriented proposals through
 time-consuming schemes. This diminishes the detectors' speed, thereby becoming
the computational bottleneck in advanced oriented object detection systems. This
 work proposes an effective and simple oriented object detection framework, term
ed Oriented R-CNN, which is a general two-stage oriented detector with promising
 accuracy and efficiency. To be specific, in the first stage, we propose an orie
nted Region Proposal Network (oriented RPN) that directly generates high-quality
 oriented proposals in a nearly cost-free manner. The second stage is oriented R
-CNN head for refining oriented Regions of Interest (oriented RoIs) and recogniz
ing them. Without tricks, oriented R-CNN with ResNet50 achieves state-of-the-art

detection accuracy on two commonly-used datasets for oriented object detection including DOTA (75.87% mAP) and HRSC2016 (96.50% mAP), while having a speed of 15.1 FPS with the image size of 1024x1024 on a single RTX 2080Ti. We hope our work could inspire rethinking the design of oriented detectors and serve as a baseline for oriented object detection. Code is available at https: //github.com/jbwang1997/OBBDetection.

****

TransVG: End-to-End Visual Grounding With Transformers

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1769-1779

In this paper, we present a neat yet effective transformer-based framework for visual grounding, namely TransVG, to address the task of grounding a language query to the corresponding region onto an image. The state-of-the-art methods, including two-stage or one-stage ones, rely on a complex module with manually-designed mechanisms to perform the query reasoning and multi-modal fusion. However, the involvement of certain mechanisms in fusion module design, such as query decomposition and image scene graph, makes the models easily overfit to datasets with specific scenarios, and limits the plenitudinous interaction between the visual-linguistic context. To avoid this caveat, we propose to establish the multi-modal correspondence by leveraging transformers, and empirically show that the complex fusion modules (e.g., modular attention network, dynamic graph, and multi-modal tree) can be replaced by a simple stack of transformer encoder layers with higher performance. Moreover, we re-formulate the visual grounding as a direct coordinates regression problem and avoid making predictions out of a set of candidates (i.e., region proposals or anchor boxes). Extensive experiments are conducted on five widely used datasets, and a series of state-of-the-art records are set by our TransVG. We build the benchmark of transformer-based visual grounding framework and make the code available at https://github.com/djiajunustc/TransVG.

****

Airbert: In-Domain Pretraining for Vision-and-Language Navigation

Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1634-1643

Vision-and-language navigation (VLN) aims to enable embodied agents to navigate in realistic environments using natural language instructions. Given the scarcity of domain-specific training data and the high diversity of image and language inputs, the generalization of VLN agents to unseen environments remains challenging. Recent methods explore pretraining to improve generalization, however, the use of generic image-caption datasets or existing small-scale VLN environments is suboptimal and results in limited improvements. In this work, we introduce BnB, a large-scale and diverse in-domain VLN dataset. We first collect image-caption (IC) pairs from hundreds of thousands of listings from online rental marketplaces. Using IC pairs we next propose automatic strategies to generate millions of VLN path-instruction (PI) pairs. We further propose a shuffling loss that improves the learning of temporal order inside PI pairs. We use BnB to pretrain our Airbert model that can be adapted to discriminative and generative settings and show that it outperforms state of the art for Room-to-Room (R2R) navigation and Remote Referring Expression (REVERIE) benchmarks. Moreover, our in-domain pretraining significantly increases performance on a challenging few-shot VLN evaluation, where we train the model only on VLN instructions from a few houses.

****

Internal Video Inpainting by Implicit Long-Range Propagation

Hao Ouyang, Tengfei Wang, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14579-14588

We propose a novel framework for video inpainting by adopting an internal learning strategy. Unlike previous methods that use optical flow for cross-frame context propagation to inpaint unknown regions, we show that this can be achieved implicitly by fitting a convolutional neural network to known regions. Moreover, to handle challenging sequences with ambiguous backgrounds or long-term occlusion,

we design two regularization terms to preserve high-frequency details and long-term temporal consistency. Extensive experiments on the DAVIS dataset demonstrate that the proposed method achieves state-of-the-art inpainting quality quantitatively and qualitatively. We further extend the proposed method to another challenging task: learning to remove an object from a video giving a single object mask in only one frame in a 4K video.
****