Towards Contextual Learning in Few-Shot Object Classification
Mathieu Page Fortin, Brahim Chaib-draa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3279-3288
Few-shot Learning (FSL) aims to classify new concepts from a small number of examples. While there have been an increasing amount of work on few-shot object classification in the last few years, most current approaches are limited to images with only one centered object. On the opposite, humans are able to leverage prior knowledge to quickly learn new concepts, such as semantic relations with contextual elements. Inspired by the concept of contextual learning in educational sciences, we propose to make a step towards adopting this principle in FSL by studying the contribution that context can have in object classification in a low-data regime. To this end, we first propose an approach to perform FSL on images of complex scenes. We develop two plug-and-play modules that can be incorporated into existing FSL methods to enable them to leverage contextual learning. More specifically, these modules are trained to weight the most important context elements while learning a particular concept, and then use this knowledge to ground visual class representations in context semantics. Extensive experiments on Visual Genome and Open Images show the superiority of contextual learning over learning individual objects in isolation.
****
Multimodal Humor Dataset: Predicting Laughter Tracks for Sitcoms
Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 576-585
A great number of situational comedies (sitcoms) are being regularly made and the task of adding laughter tracks to these is a critical task. Providing an ability to be able to predict whether something will be humorous to the audience is also crucial. In this project, we aim to automate this task. Towards doing so, we annotate an existing sitcom (`Big Bang Theory') and use the laughter cues present to obtain a manual annotation for this show. We provide detailed analysis for the dataset design and further evaluate various state of the art baselines for solving this task. We observe that existing LSTM and BERT based networks on the text alone do not perform as well as joint text and video or only video-based networks. Moreover, it is challenging to ascertain that the words attended to while predicting laughter are indeed humorous. Our dataset and analysis provided through this paper is a valuable resource towards solving this interesting semantic and practical task. As an additional contribution, we have developed a novel model for solving this task that is a multi-modal self-attention based model that outperforms currently prevalent models for solving this task. The project page for the submission is \url https://multimodal-humor-dataset.github.io/
****
Self-Supervised Learning for Domain Adaptation on Point Clouds
Idan Achituve, Haggai Maron, Gal Chechik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 123-133
Self-supervised learning (SSL) is a technique for learning useful representations from unlabeled data. It has been applied effectively to domain adaptation (DA) on images and videos. It is still unknown if and how it can be leveraged for domain adaptation in 3D perception problems. Here we describe the first study of SSL for DA on point clouds. We introduce a new family of pretext tasks, Deformation Reconstruction, inspired by the deformations encountered in sim-to-real transformations. In addition, we propose a novel training procedure for labeled point cloud data motivated by the MixUp method called Point cloud Mixup (PCM). Evaluations on domain adaptations datasets for classification and segmentation, demonstrate a large improvement over existing and baseline methods.
****
Efficient 3D Video Engine Using Frame Redundancy
Gao Peng, Bo Pang, Cewu Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3792-3802
Traditional 3d video understanding methods process videos frame by frame. We argue that a lot of computation in this mechanism is redundant based on a key obser

vation - adjacent frames in 3D videos have visually similar geometry structure. To handle the redundancy, we propose the Efficient 3D Video Engine (EVE), aiming to avoid the computation of redundant points. It consists of two modules: 1) redundancy removing module designed to detect redundancy and remove it; 2) residual learning module to extract features on non-redundant points. As a simple plug and play framework, EVE can be easily incorporated in main-stream 3D models. Experiments demonstrate that EVE can significantly reduce computation without performance loss on large scale datasets. On the other hand, with similar computation, EVE outperforms the strong baseline by up to 4.1 mIoU on SemanticKITTI. The code is available on https://github.com/ecr23xx/eve.

****

Exploiting the Redundancy in Convolutional Filters for Parameter Reduction

Kumara Kahatapitiya, Ranga Rodrigo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1410-1420

Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in many computer vision tasks over the years. However, this comes at the cost of heavy computation and memory intensive network designs, suggesting potential improvements in efficiency. Convolutional layers of CNNs partly account for such an inefficiency, as they are known to learn redundant features. In this work, we exploit this redundancy, observing it as the correlation between convolutional filters of a layer, and propose an alternative approach to reproduce it efficiently. The proposed 'LinearConv' layer learns a set of orthogonal filters, and a set of coefficients that linearly combines them to introduce a controlled redundancy. We introduce a correlation-based regularization loss to achieve such flexibility over redundancy, and control the number of parameters in turn. This is designed as a plug-and-play layer to conveniently replace a conventional convolutional layer, without any additional changes required in the network architecture or the hyperparameter settings. Our experiments verify that LinearConv models achieve a performance on-par with their counterparts, with almost a 50% reduction in parameters on average, and the same computational requirement and speed at inference. Source is available at https://github.com/kkahatapitiya/LinearConv .

****

Benchmark for Evaluating Pedestrian Action Prediction

Iuliia Kotseruba, Amir Rasouli, John K. Tsotsos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1258-1268

Pedestrian action prediction has been a topic of active research in recent years resulting in many new algorithmic solutions. However, measuring the overall progress towards solving this problem is difficult due to the lack of publicly available benchmarks and common training and evaluation procedures. To this end, we introduce a benchmark based on two public datasets for pedestrian behavior understanding. Using the proposed evaluation procedures, we rank a number of baseline and state-of-the-art models and analyze their performance with respect to various properties of the data. Based on these findings we propose a new model for pedestrian crossing action prediction that uses attention mechanisms to effectively combine implicit and explicit features and demonstrate new state-of-the-art results. The code for models and evaluation is available at https://github.com/ykotseruba/ PedestrianActionBenchmark.

****

Regional Attention Networks With Context-Aware Fusion for Group Emotion Recognition

Ahmed Shehab Khan, Zhiyuan Li, Jie Cai, Yan Tong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1150-1159

Group Emotion Recognition (GER) from images has many inherent challenges. Specifically, it is difficult to combine diverse emotions of different individuals into a single conclusive label. In addition, although utilization of information other than faces like scene and objects has proven helpful, it is still a challenge to effectively fuse predictions of individual sources. In this work, we proposed solutions to these two problems. First, we developed a regional attention mechanism to find important persons or objects, which play critical roles in the group emotion, and combine them based on importance. Second, we proposed a context

-aware fusion mechanism to estimate weights from the image context to fuse different sources of information. Finally, we proposed to use a single backbone network to extract features from multiple sources, i.e., scene, faces, and objects, cutting down computation and memory cost. Experiments on two GER datasets have shown that the proposed framework achieves performance comparable to the state-of-the-art. Furthermore, a visualization study and a case study have demonstrated that the proposed model is effective to extract and more importantly, emphasize the most critical information in GER.

****

Auxiliary Tasks for Efficient Learning of Point-Goal Navigation

Saurabh Satish Desai, Stefan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 717-725

Top-performing approaches to embodied AI tasks like point-goal navigation often rely on training agents via reinforcement learning over tens of millions (or even billions) of experiential steps -- learning neural agents that map directly from visual observations to actions. In this work, we question whether these extreme training durations are necessary or if they are simply due to the difficulty of learning visual representations purely from task reward. We examine the task of point-goal navigation in photorealistic environments and introduce three auxiliary tasks that encourage learned representations to capture key elements of the task -- local scene geometry, transition dynamics of the environment, and progress towards the goal. Importantly, these can be evaluated independent of task performance and provide strong supervision for representation learning. Our auxiliary tasks are simple to implement and rely on supervision already present in simulators commonly used for point-goal navigation. Applying our auxiliary losses to agents from prior works, we observe a greater than 4 times improvement in sample efficiency -- in 17 million steps, our augmented agents outperform state-of-the-art agents after 72 million steps.

****

Cross-Modality 3D Object Detection

Ming Zhu, Chao Ma, Pan Ji, Xiaokang Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3772-3781

In this paper, we focus on exploring the fusion of images and point clouds for 3D object detection in view of the complementary nature of the two modalities, i.e., images possess more semantic information while point clouds specialize in distance sensing. To this end, we present a novel two-stage multi-modal fusion network for 3D object detection, taking both binocular images and raw point clouds as input. The whole architecture facilitates two-stage fusion. The first stage aims at producing 3D proposals through point-wise feature fusion. Within the first stage, we further exploit a joint anchor mechanism that enables the network to utilize 2D-3D classification and regression simultaneously for better proposal generation. The second stage works on the 2D and 3D proposal regions and fuses their features. In addition, we propose to use pseudo LiDAR points from stereo matching as a data augmentation method to densify the LiDAR points, as we observe that objects missed by the detection network mostly have too few points, especially for far-away objects. Our experiments on the KITTI dataset show that the proposed multi-stage fusion helps the network to learn better representations.

****

Variational Prototype Inference for Few-Shot Semantic Segmentation

Haochen Wang, Yandan Yang, Xianbin Cao, Xiantong Zhen, Cees Snoek, Ling Shao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 525-534

In this paper, we propose variational prototype inference to address few-shot semantic segmentation in a probabilistic framework. A probabilistic latent variable model infers the distribution of the prototype that is treated as the latent variable. We formulate the optimization as a variational inference problem, which is established with an amortized inference network based on an auto-encoder architecture. The probabilistic modeling of the prototype enhances its generalization ability to handle the inherent uncertainty caused by limited data and the huge intra-class variations of objects. Moreover, it offers a principled way to inc orporate the prototype extracted from support images into the prediction of the segmentation maps for query images. We conduct extensive experimental evaluation on three benchmark datasets. Ablation studies show the effectiveness of variational prototype inference for few-shot semantic segmentation by probabilistic modeling. On all three benchmarks, our proposal achieves high segmentation accuracy and surpasses previous methods by considerable margins.

****

Zero-Shot Recognition via Optimal Transport

Wenlin Wang, Hongteng Xu, Guoyin Wang, Wenqi Wang, Lawrence Carin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3471-3481

We propose an optimal transport (OT) framework for generalized zero-shot learning (GZSL), seeking to distinguish samples for both seen and unseen classes, with the assist of auxiliary attributes. The discrepancy between features and attributes is minimized by solving an optimal transport problem. Specifically, we build a conditional generative model to generate features from seen-class attributes, and establish an optimal transport between the distribution of the generated features and that of the real features. The generative model and the optimal transport are optimized iteratively with an attribute-based regularizer, that further enhances the discriminative power of the generated features. A classifier is learned based on the features generated for both the seen and unseen classes. In addition to generalized zero-shot learning, our framework is also applicable to standard and transductive ZSL problems. Experiments show that our optimal transport-based method outperforms state-of-the-art methods on several benchmark datasets.

****

Multi-Modal Reasoning Graph for Scene-Text Based Fine-Grained Image Classification and Retrieval

Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4023-4033

Scene text instances found in natural images carry explicit semantic information that can provide important cues to solve a wide array of computer vision problems. In this paper, we focus on leveraging multi-modal content in the form of visual and textual cues to tackle the task of fine-grained image classification and retrieval. First, we obtain the text instances from images by employing a text reading system. Then, we combine textual features with salient image regions to exploit the complementary information carried by the two sources. Specifically, we employ a Graph Convolutional Network to perform multi-modal reasoning and obtain relationship-enhanced features by learning a common semantic space between salient objects and text found in an image. By obtaining an enhanced set of visual and textual features, the proposed model greatly outperforms previous state-of-the-art in two different tasks, fine-grained classification and image retrieval in the Con-Text and Drink Bottle datasets.

****

PI-Net: Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation

Wen Guo, Enric Corona, Francesc Moreno-Noguer, Xavier Alameda-Pineda; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2796-2806

Recent literature addressed the monocular 3D pose estimation task very satisfactorily. In these studies, different persons are usually treated as independent pose instances to estimate. However, in many every-day situations, people are interacting, and the pose of an individual depends on the pose of his/her interactees. In this paper, we investigate how to exploit these dependency to enhance current, and possibly future, deep networks for 3D monocular pose estimation. Our pose interacting network, or PI-Net, inputs the initial pose estimates of a variable number of interactees into a recurrent network used to refine the pose of the person-of-interest. Evaluating such a method is challenging due to the limited availability of public annotated multi-person 3D human pose datasets. We demonstrate the effectiveness of our method in the MuPoTS dataset, setting the new stat e-of-the-art on it. Qualitative results on other multi-person datasets (for which 3D pose ground-truth is not available) showcase the proposed PI-Net.

****

Size-Invariant Detection of Marine Vessels From Visual Time Series
Tunai Porto Marques, Alexandra Branzan Albu, Patrick O'Hara, Norma Serra, Ben Morrow, Lauren McWhinnie, Rosaline Canessa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 443-453

Marine vessel traffic is one of the main sources of negative anthropogenic impact upon marine environments. The automatic identification of boats in monitoring images facilitates conservation, research and patrolling efforts. However, the diverse sizes of vessels, the highly dynamic water surface and weather-related visibility issues significantly hinder this task. While recent deep learning (DL)-based object detectors identify well medium- and large-sized boats, smaller vessels, often responsible for substantial disturbance to sensitive marine life, are typically not detected. We propose a detection approach that combines state-of-the-art object detectors and a novel Detector of Small Marine Vessels (DSMV) to identify boats of any size. The DSMV uses a short time series of images and a novel bi-directional Gaussian Mixture technique to determine motion in combination with context-based filtering and a DL-based image classifier. Experimental results obtained on our publicly-released datasets of images containing boats of various sizes show that the proposed approach comfortably outperforms five popular state-of-the-art object detectors. Code and datasets available at https://github.com/tunai/hybrid-boat-detection.

****

Temporal Stochastic Softmax for 3D CNNs: An Application in Facial Expression Recognition
Theo Ayral, Marco Pedersoli, Simon Bacon, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3029-3038

Training deep learning models for accurate spatiotemporal recognition of facial expressions in videos requires significant computational resources. For practical reasons, 3D Convolutional Neural Networks (3D CNNs) are usually trained with relatively short clips randomly extracted from videos. However, such uniform sampling is generally sub-optimal because equal importance is assigned to each temporal clip. In this paper, we present a strategy for efficient video-based training of 3D CNNs. It relies on softmax temporal pooling and a weighted sampling mechanism to select the most relevant training clips. The proposed softmax strategy provides several advantages - a reduced computational complexity due to efficient clip sampling, and an improved accuracy since temporal weighting focuses on more relevant clips during both training and inference. Experimental results obtained with the proposed method on several facial expression recognition benchmarks show the benefits of focusing on more informative clips in training videos. In particular, our approach improves performance and computational cost by reducing the impact of inaccurate trimming and coarse annotation of videos, and heterogeneous distribution of visual information across time.

****

Scale Aware Adaptation for Land-Cover Classification in Remote Sensing Imagery
Xueqing Deng, Yi Zhu, Yuxin Tian, Shawn Newsam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2160-2169

Land-cover classification using remote sensing imagery is an important Earth observation task. Recently, land cover classification has benefited from the development of fully connected neural networks for semantic segmentation. The benchmark datasets available for training deep segmentation models in remote sensing imagery tend to be small, however, often consisting of only a handful of images from a single location with a single scale. This limits the models' ability to generalize to other datasets. Domain adaptation has been proposed to improve the models' generalization but we find these approaches are not effective for dealing with the scale variation commonly found between remote sensing image collections. We therefore propose a scale aware adversarial learning framework to perform joint cross-location and cross-scale land-cover classification. The framework has a dual discriminator architecture with a standard feature discriminator as well as a novel scale discriminator. We also introduce a scale attention module which produces scale-enhanced features. Experimental results show that the proposed framework outperforms state-of-the-art domain adaptation methods by a large margin. The open-sourced codes are available on Github: https://github.com/xdeng7/scale-aware_da.

****

Intro and Recap Detection for Movies and TV Series

Xiang Hao, Kripa Chettiar, Ben Cheung, Vernon Germano, Raffay Hamid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 167-176

Modern video streaming service companies offer millions of video-titles for its customers. A lot of these titles have repetitive introductory and recap parts in the beginning that customers have to manually skip in order to achieve an uninterrupted viewing experience. To avoid this unnecessary friction, some of the services have recently added "skip-intro" and "skip-recap" buttons to their video players before the intro and recap parts start. To efficiently scale this experience to their entire catalogs, it is important to automate the process of finding the intro and recap portions of titles. In this work, we pose intro and recap detection as a supervised sequence labeling problem and propose a novel end-to-end deep learning framework to this end. Specifically, we use CNNs to extract both visual and audio features from videos, and fuse these features using a B-LSTM in order to capture the various long and short term dependencies among different frame-features over time. Finally, we use a CRF to jointly optimize the sequence labeling for the intro and recap parts of the titles. We present a thorough empirical analysis of our model compared to several other deep learning based architectures and demonstrate the superior performance of our approach.

****

Supervoxel Attention Graphs for Long-Range Video Modeling

Yang Wang, Gedas Bertasius, Tae-Hyun Oh, Abhinav Gupta, Minh Hoai, Lorenzo Torresani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 155-166

A significant challenge in video understanding is posed by the high dimensionality of the input, which induces large computational cost and high memory footprints. Deep convolutional models operating on video apply pooling and striding to reduce feature dimensionality and to increase the receptive field. However, despite these strategies, modern approaches cannot effectively leverage spatiotemporal structure over long temporal extents. In this paper we introduce an approach that reduces a video of 10 seconds to a sparse graph of only 160 feature nodes such that efficient inference in this graph produces state-of-the-art accuracy on challenging action recognition datasets. The nodes of our graph are semantic supervoxels that capture the spatiotemporal structure of objects and motion cues in the video, while edges between nodes encode spatiotemporal relations and feature similarity. We demonstrate that a shallow network that interleaves graph convolution and graph pooling on this compact representation implements an effective mechanism of relational reasoning yielding strong recognition results on both Charades and Something-Something.

****

SoFA: Source-Data-Free Feature Alignment for Unsupervised Domain Adaptation

Hao-Wei Yeh, Baoyao Yang, Pong C. Yuen, Tatsuya Harada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 474-483

Applying a trained model on a new scenario may suffer from domain shift. Unsupervised domain adaptation (UDA) has been proven to be an effective approach to solve the problem of domain shift by leveraging both data from the scenario that the model was trained on (source) and the new scenario (target). Although the source data are available for training the source model, there is no guarantee that the source data will still be available when applying UDA in the future due to emerging regulations on privacy of data. This results in the in-applicability of most existing UDA methods in the absence of source data. This paper proposes a s ource-data-free feature alignment (SoFA) method to address this problem by only using the trained source model and unlabeled target data. The source model is used to predict the labels for target data, and we model the generation process from predicted classes to input data to infer the latent features for alignment. Specifically, a mixture of Gaussian distributions is induced from the predicted classes as the reference distribution. The encoded target features are then aligned to the reference distribution via variational inference to extract class semantics without accessing source data. Relationship of the proposed method and the theory of domain adaptation is provided to verify the performance. Experimental results show the proposed method achieves higher or comparable accuracy compared to the existing methods in several cross-dataset classification tasks. Ablation studies are also conducted to confirm the importance of latent feature alignment to adaptation performance.

****

Towards Visually Explaining Video Understanding Networks With Perturbation
Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, Yoichi Sato; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1120-1129

'Making black box models explainable' is a vital problem that accompanies the development of deep learning networks. For networks taking visual information as input, one basic but challenging explanation method is to identify and visualize the input pixels/regions that dominate the network's prediction. However, most existing works focus on explaining networks taking a single image as input and do not consider the temporal relationship that exists in videos. Providing an easy-to-use visual explanation method that is applicable to diversified structures of video understanding networks still remains an open challenge. In this paper, we investigate a generic perturbation-based method for visually explaining video understanding networks. Besides, we propose a novel loss function to enhance the method by constraining the smoothness of its results in both spatial and temporal dimensions. The method enables the comparison of explanation results between different network structures to become possible and can also avoid generating the pathological adversarial explanations for video inputs. Experimental comparison results verified the effectiveness of our method.

****

Continual Representation Learning for Biometric Identification
Bo Zhao, Shixiang Tang, Dapeng Chen, Hakan Bilen, Rui Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1198-1208

With the explosion of digital data in recent years, continuously learning new tasks from a stream of data without forgetting previously acquired knowledge has become increasingly important. In this paper, we propose a new continual learning (CL) setting, namely "continual representation learning", which focuses on learning better representation in a continuous way. We also provide two large-scale multi-step benchmarks for biometric identification, where the visual appearance of different classes are highly relevant. In contrast to requiring the model to recognize more learned classes, we aim to learn feature representation that can be better generalized to not only previously unseen images but also unseen classes/identities. For the new setting, we propose a novel approach that performs the knowledge distillation over a large number of identities by applying the neighbourhood selection and consistency relaxation strategies to improve scalability and flexibility of the continual learning model. We demonstrate that existing CL methods can improve the representation in the new setting, and our method achieves better results than the competitors.

****

MUSCLE: Strengthening Semi-Supervised Learning via Concurrent Unsupervised Learning Using Mutual Information Maximization
Hanchen Xie, Mohamed E. Hussein, Aram Galstyan, Wael Abd-Almageed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2586-2595

Deep neural networks are powerful, massively parameterized machine learning mode ls that have been shown to perform well in supervised learning tasks. However, very large amounts of labeled data are usually needed to train deep neural networks. Several semi-supervised learning approaches have been proposed to train neural networks using smaller amounts of labeled data with a large amount of unlabeled data. The performance of these semi-supervised methods significantly degrades as the size of labeled data decreases. We introduce Mutual-information-based Unsupervised & Semi-supervised Concurrent LEarning (MUSCLE), a hybrid learning approach that uses mutual information to combine both unsupervised and semi-supervised learning. MUSCLE can be used as a stand-alone training scheme for neural networks, and can also be incorporated into other learning approaches. We show that the proposed hybrid model outperforms state of the art on several standard benchmarks, including CIFAR-10, CIFAR-100, and Mini-Imagenet. Furthermore, the performance gain consistently increases with the reduction in the amount of labeled data, as well as in the presence of bias. We also show that MUSCLE has the potential to boost the classification performance when used in the fine-tuning phase for a model pre-trained only on unlabeled data.

****

AVGZSLNet: Audio-Visual Generalized Zero-Shot Learning by Reconstructing Label Features From Multi-Modal Embeddings

Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3090-3099

In this paper, we propose a novel approach for generalized zero-shot learning in a multi-modal setting, where we have novel classes of audio/video during testing that are not seen during training. We use the semantic relatedness of text embeddings as a means for zero-shot learning by aligning audio and video embeddings with the corresponding class label text feature space. Our approach uses a cross-modal decoder and a composite triplet loss. The cross-modal decoder enforces a constraint that the class label text features can be reconstructed from the audio and video embeddings of data points. This helps the audio and video embeddings to move closer to the class label text embedding. The composite triplet loss makes use of the audio, video, and text embeddings. It helps bring the embeddings from the same class closer and push away the embeddings from different classes in a multi-modal setting. This helps the network to perform better on the multi-modal zero-shot learning task. Importantly, our multi-modal zero-shot learning approach works even if a modality is missing at test time. We test our approach on the generalized zero-shot classification and retrieval tasks and show that our approach outperforms other models in the presence of a single modality as well as in the presence of multiple modalities. We validate our approach by comparing it with previous approaches and using various ablations.

****

Continuous Geodesic Convolutions for Learning on 3D Shapes

Zhangsihao Yang, Or Litany, Tolga Birdal, Srinath Sridhar, Leonidas Guibas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 134-144

The majority of descriptor-based methods for geometric processing of non-rigid shape rely on hand-crafted descriptors. Recently, learning-based techniques have been shown effective, achieving state-of-the-art results in a variety of tasks. Yet, even though these methods can in principle work directly on raw data, most methods still rely on hand-crafted descriptors at the input layer. In this work, we wish to challenge this practice and use a neural network to learn descriptors directly from the raw mesh. To this end, we introduce two modules into our neural architecture. The first is a local reference frame (LRF) used to explicitly make the features invariant to rigid transformations. The second is continuous convolution kernels that provide robustness to sampling. We show the efficacy of our proposed network in learning on raw meshes using two cornerstone tasks: shape matching, and human body parts segmentation. Our results show superior results over baseline methods that use hand-crafted descriptors.

****

Shape From Caustics: Reconstruction of 3D-Printed Glass From Simulated Caustic I mages

Marc Kassubeck, Florian Burgel, Susana Castillo, Sebastian Stiller, Marcus Magnor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2877-2886

We present an efficient and effective computational framework for the inverse rendering problem of reconstructing the 3D shape of a piece of glass from its caustic image. Our approach is motivated by the needs of 3D glass printing, a nascent additive manufacturing technique that promises to revolutionize the production of optics elements, from lightweight mirrors to waveguides and lenses. One important problem is the reliable control of the manufacturing process by inferring the printed 3D glass shape from its caustic image. Towards this goal, we propose a novel general-purpose reconstruction algorithm based on differentiable light propagation simulation followed by a regularization scheme that takes the deposited glass volume into account. This enables incorporating arbitrary measurements of caustics into an efficient reconstruction framework. We demonstrate the effectiveness of our method and establish the influence of our hyperparameters using several sample shapes and parameter configurations.

****

Weakly Supervised Instance Segmentation by Deep Community Learning

Jaedong Hwang, Seohyun Kim, Jeany Son, Bohyung Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1020-1029

We present a weakly supervised instance segmentation algorithm based on deep community learning with multiple tasks. This task is formulated as a combination of weakly supervised object detection and semantic segmentation, where individual objects of the same class are identified and segmented separately. We address this problem by designing a unified deep neural network architecture, which has a positive feedback loop of object detection with bounding box regression, instance mask generation, instance segmentation, and feature extraction. Each component of the network makes active interactions with others to improve accuracy, and the end-to-end trainability of our model makes our results more robust and reproducible. The proposed algorithm achieves state-of-the-art performance in the weakly supervised setting without any additional training such as Fast R-CNN and Mask R-CNN on the standard benchmark dataset. The implementation of our algorithm is available on the project webpage: https://cv.snu.ac.kr/research/WSIS_CL.

****

Coarse-to-Fine Gaze Redirection With Numerical and Pictorial Guidance

Jingjing Chen, Jichao Zhang, Enver Sangineto, Tao Chen, Jiayuan Fan, Nicu Sebe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3665-3674

Gaze redirection aims at manipulating the gaze of a given face image with respect to a desired direction (i.e., a reference angle) and it can be applied to many real life scenarios, such as video-conferencing or taking group photos. However, previous work on this topic mainly suffers of two limitations: (1) Low-quality image generation and (2) Low redirection precision. In this paper, we propose to alleviate these problems by means of a novel gaze redirection framework which exploits both a numerical and a pictorial direction guidance, jointly with a coarse-to-fine learning strategy. Specifically, the coarse branch learns the spatial transformation which warps input image according to desired gaze. On the other hand, the fine-grained branch consists of a generator network with conditional residual image learning and a multi-task discriminator. This second branch reduces the gap between the previously warped image and the ground-truth image and recovers finer texture details. Moreover, we propose a numerical and pictorial guidance module (NPG) which uses a pictorial gazemap description and numerical angles as an extra guide to further improve the precision of gaze redirection. Extensive experiments on a benchmark dataset show that the proposed method outperforms the state-of-the-art approaches in terms of both image quality and redirection precision.

****

S3-Net: A Fast and Lightweight Video Scene Understanding Network by Single-Shot Segmentation

Yuan Cheng, Yuchao Yang, Hai-Bao Chen, Ngai Wong, Hao Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3329-3337

Real-time understanding in video is crucial in various AI applications such as autonomous driving. This work presents a fast single-shot segmentation strategy for video scene understanding. The proposed net, called S3-Net, quickly locates and segments target sub-scenes, meanwhile extracts structured time-series semantic features as inputs to an LSTM-based spatio-temporal model. Utilizing tensorization and quantization techniques, S3-Net is intended to be lightweight for edge computing. Experiments using CityScapes, UCF11, HMDB51 and MOMENTS datasets demonstrate that the proposed S3-Net achieves an accuracy improvement of 8.1% versus the 3D-CNN based approach on UCF11, a storage reduction of 6.9x and an inference speed of 22.8 FPS on CityScapes with a GTX1080Ti GPU.

****

Improve CAM With Auto-Adapted Segmentation and Co-Supervised Augmentation

Ziyi Kou, Guofeng Cui, Shaojie Wang, Wentian Zhao, Chenliang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3598-3606

Weakly Supervised Object Localization (WSOL) methods generate both classification and localization results by learning from only image category labels. Previous methods usually utilize class activation map (CAM) to obtain target object regions. However, most of them only focus on improving foreground object parts in CAM, but ignore the important effect of its background contents. In this paper, we propose a confidence segmentation (ConfSeg) module that builds a confidence score for each pixel in CAM without introducing additional hyper-parameters. The generated sample-specific confidence mask is able to indicate the extent of determination for each pixel in CAM, and further supervises additional CAM extended from internal feature maps. Besides, we introduce Co-supervised Augmentation (CoAug) module to capture feature-level representation for foreground and background parts in CAM separately. Then a metric loss is applied at batch sample level to augment distinguish ability of our model, which helps a lot to localize more related object parts. Our final model, CSoA, combines the two modules and achieves superior performance, e.g. 37.69% and 48.81% Top-1 localization error on CUB-200 and ILSVRC datasets, respectively, which outperforms all previous methods and becomes the new state-of-the-art.

****

Analyzing Deep Neural Network's Transferability via Frechet Distance

Yifan Ding, Liqiang Wang, Boqing Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3932-3941

Transfer learning has become the de facto practice to reuse a deep neural network (DNN) that is pre-trained with abundant training data in a source task to improve the model training on target tasks with smaller-scale training data. In this paper, we first investigate the correlation between the DNN's pre-training performance in the source task and their transfer results in the downstream tasks. We find that high performance of a pre-trained model does not necessarily imply high transferability. We then propose a metric, named Fr echet Pre-train Distance, to estimate the transferability of a deep neural network. By applying the proposed Fr echet Pre-train Distance, we are able to identify the optimal pre-trained checkpoint, and then achieve high transferability on downstream tasks. Finally, we investigate several factors impacting DNN's transferability including normalization, different networks and learning rates. The results consistently support our conclusions.

****

Controllable and Progressive Image Extrapolation

Yijun Li, Lu Jiang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2140-2149

Image extrapolation aims at expanding the narrow field of view of a given image patch. Existing models mainly deal with natural scene images of homogeneous regions and have no control of the content generation process. In this work, we study conditional image extrapolation to synthesize new images guided by the input s tructured text. The text is represented as a graph to specify the objects and their spatial relation to the unknown regions of the image. Inspired by drawing techniques, we propose a progressive generative model of three stages, i.e., generating a coarse bounding-boxes layout, refining it to a finer segmentation layout, and mapping the layout to a realistic output. Such a multi-stage design is shown to facilitate the training process and generate more controllable results. We validate the effectiveness of the proposed method on the face and human clothing dataset in terms of visual results, quantitative evaluations, and flexible controls.

****

How to Make a BLT Sandwich? Learning VQA Towards Understanding Web Instructional Videos

Shaojie Wang, Wentian Zhao, Ziyi Kou, Jing Shi, Chenliang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1130-1139

Understanding web instructional videos is an essential branch of video understanding in two aspects. First, most existing video methods focus on short-term actions for a-few-second-long video clips; these methods are not directly applicable to long videos. Second, unlike unconstrained long videos, e.g., movies, instructional videos are more structured in that they have step-by-step procedures constraining the understanding task. In this work, we study problem-solving on instructional videos via Visual Question Answering (VQA). Surprisingly, it has not been an emphasis for the video community despite its rich applications. We thereby introduce YouCookQA, an annotated QA dataset for instructional videos based on YouCook2. The questions in YouCookQA are not limited to cues on a single frame but relations among multiple frames in the temporal dimension. Observing the lack of effective representations for modeling long videos, we propose a set of carefully designed models including a Recurrent Graph Convolutional Network (RGCN) that captures both temporal order and relational information. Furthermore, we study multiple modalities including descriptions and transcripts for the purpose of boosting video understanding. Extensive experiments on YouCookQA suggest that RGCN performs the best in terms of QA accuracy and better performance is gained by introducing human-annotated descriptions. YouCookQA dataset is available at https://github.com/Jossome/YoucookQA.

****

Proposal Learning for Semi-Supervised Object Detection

Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, Caiming Xiong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2291-2301

In this paper, we focus on semi-supervised object detection to boost performance of proposal-based object detectors (a.k.a. two-stage object detectors) by training on both labeled and unlabeled data. However, it is non-trivial to train object detectors on unlabeled data due to the unavailability of ground truth labels. To address this problem, we present a proposal learning approach to learn proposal features and predictions from both labeled and unlabeled data. The approach consists of a self-supervised proposal learning module and a consistency-based proposal learning module. In the self-supervised proposal learning module, we present a proposal location loss and a contrastive loss to learn context-aware and noise-robust proposal features respectively. In the consistency-based proposal learning module, we apply consistency losses to both bounding box classification and regression predictions of proposals to learn noise-robust proposal features and predictions. Our approach enjoys the following benefits: 1) encouraging more context information to be delivered in the proposals learning procedure; 2) noisy proposal features and enforcing consistency to allow noise-robust object detection; 3) building a general and high-performance semi-supervised object detection framework, which can be easily adapted to proposal-based object detectors with different backbone architectures. Experiments are conducted on the COCO dataset with all available labeled and unlabeled data. Results demonstrate that our approach consistently improves the performance of fully-supervised baselines. In particular, after combining with data distillation [38], our approach improves AP by about 2.0% and 0.9% on average compared to fully-supervised baselines and data distillation baselines respectively.

****

Multi-Path Neural Networks for On-Device Multi-Domain Visual Classification

Qifei Wang, Junjie Ke, Joshua Greaves, Grace Chu, Gabriel Bender, Luciano Sbaiz, Alec Go, Andrew Howard, Ming-Hsuan Yang, Jeff Gilbert, Peyman Milanfar, Feng Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3019-3028

Learning multiple domains/tasks with a single model is important for improving data efficiency and lowering inference cost for numerous vision tasks, especially on resource-constrained mobile devices. However, hand-crafting a multi-domain/task model can be both tedious and challenging. This paper proposes a novel approach to automatically learn a multi-path network for multi-domain visual classification on mobile devices. The proposed multi-path network is learned from neural architecture search by applying one reinforcement learning controller for each domain to select the best path in the super-network created from a MobileNetV3-like search space. An adaptive balanced domain prioritization algorithm is proposed to balance optimizing the joint model on multiple domains simultaneously. The determined multi-path model selectively shares parameters across domains in shared nodes while keeping domain-specific parameters within non-shared nodes in individual domain paths. This approach effectively reduces the total number of parameters and FLOPS, encouraging positive knowledge transfer while mitigating negative interference across domains. Extensive evaluations on the Visual Decathlon dataset demonstrate that the proposed multi-path model achieves state-of-the-art performance in terms of accuracy, model size, and FLOPS against other approaches using MobileNetV3-like architectures. Furthermore, the proposed method improves average accuracy over learning single-domain models individually, and reduces the total number of parameters and FLOPS by 78% and 32% respectively, compared to the approach that simply bundles single-domain models for multi-domain learning.

****

Effectiveness of Arbitrary Transfer Sets for Data-Free Knowledge Distillation

Gaurav Kumar Nayak, Konda Reddy Mopuri, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1430-1438

Knowledge Distillation is an effective method to transfer the learning across deep neural networks. Typically, the dataset originally used for training the Teacher model is chosen as the "Transfer Set" to conduct the knowledge transfer to the Student. However, this original training data may not always be freely available due to privacy or sensitivity concerns. In such scenarios, existing approaches either iteratively compose a synthetic set representative of the original training dataset, one sample at a time or learn a generative model to compose such a transfer set. However, both these approaches involve complex optimization (GAN training or several backpropagation steps to synthesize one sample) and are often computationally expensive. In this paper, as a simple alternative, we investigate the effectiveness of "arbitrary transfer sets" such as random noise, publicly available synthetic, and natural datasets, all of which are completely unrelated to the original training dataset in terms of their visual or semantic contents. Through extensive experiments on multiple benchmark datasets such as MNIST, FMNIST, CIFAR-10 and CIFAR-100, we discover and validate surprising effectiveness of using arbitrary data to conduct knowledge distillation when this dataset is "target-class balanced". We believe that this important observation can potentially lead to designing baselines for the data-free knowledge distillation task.

****

Relighting Images in the Wild With a Self-Supervised Siamese Auto-Encoder

Yang Liu, Alexandros Neophytou, Sunando Sengupta, Eric Sommerlade; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 32-40

We propose a self-supervised method for image relighting of single view images in the wild. The method is based on an auto-encoder which deconstructs an image i nto two separate encodings, relating to the scene illumination and content. In order to disentangle this embedding information without supervision, we exploit the assumption that some augmented operations do not affect the image content and only affect the direction of the light. A novel loss function, called spherical harmonic loss, is introduced that forces the illumination embedding to convert to a spherical harmonic vector. We train our model on large-scale data-sets such as Youtube 8M and CelebA. Our experiments show that our method can correctly estimate scene illumination and generate realistic re-lit examples, without any supervision or a prior shape model. Compared to supervised methods, our approach has similar performance and avoids common lighting artifacts.

****

Attention-Based Spatial Guidance for Image-to-Image Translation

Yu Lin, Yigong Wang, Yifan Li, Yang Gao, Zhuoyi Wang, Latifur Khan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 816-825

The aim of image-to-image translation algorithms is to tackle the challenges of learning a proper mapping function across different domains. Generative Adversarial Networks (GAN) have shown superior ability to handle this problem by both supervised and unsupervised ways. However, one critical problem of GAN in practice is that the discriminator is typically much stronger than the generator, which could lead to failures such as mode collapse, diminished gradient, etc. To address these shortcomings, we propose a novel framework, which incorporates a powerful spatial attention mechanism to guide the generator. Specifically, our designed discriminator estimates the probability of realness of a given image, and provides an attention map regarding this prediction. The generated attention map contains the informative regions to distinguish the real and fake image, from the perspective of the discriminator. Such information is particularly valuable for the translation because the generator is encouraged to focus on those areas and produce more realistic images. We conduct extensive experiments and evaluations, and show that our proposed method is both qualitatively and quantitatively better than other state-of-the-art image translation frameworks.

****

EVET: Enhancing Visual Explanations of Deep Neural Networks Using Image Transformations

Youngrock Oh, Hyungsik Jung, Jeonghyung Park, Min Soo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3579-3587

Numerous interpretability methods have been developed to visually explain the behavior of complex machine learning models by estimating parts of the input image that are critical for the model's prediction. We propose a general pipeline of enhancing visual explanations using image transformations (EVET). EVET considers transformations of the original input image to refine the critical input region based on an intuitive rationale that the region estimated to be important in variously transformed inputs is more important. Our proposed EVET is applicable to existing visual explanation methods without modification. We validate the effectiveness of the proposed method qualitatively and quantitatively to show that the resulting explanation method outperforms the original in terms of faithfulness, localization, and stability. We also demonstrate that EVET can be used to achieve desirable performance with a low computational cost. For example, EVET-applied Grad-CAM achieves performance comparable to Score-CAM, which is the state-ofthe-art activation-based explanation method, while reducing execution time by more than 90% on VOC, COCO, and ImageNet.

****

ATM: Attentional Text Matting

Peng Kang, Jianping Zhang, Chen Ma, Guiling Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3902-3911

Image matting is a fundamental computer vision problem and has many applications. Previous image matting methods always focus on extracting a general object or portrait from the background in an image. In this paper, we try to solve the text matting problem, which extracts characters (usually WordArts) from the backgro und in an image. Different from traditional image matting problems, text matting is much harder because of its foreground's three properties: smallness, multi-objectness, and complicated structures and boundaries. We propose a two-stage attentional text matting pipeline to solve the text matting problem. In the first stage, we utilize text detection methods to serve as the attention mechanism. In the second stage, we employ the attentional text regions and matting system to obtain mattes of these text regions. Finally, we post-process the mattes and obtain the final matte of the input image. We also construct a large-scale dataset with high-quality annotations consisting of 46,289 unique foregrounds to facilitate the learning and evaluation of text matting. Extensive experiments on this dataset and real images clearly demonstrate the superiority of our proposed pipeline over previous image matting methods on the task of text matting.

****

Efficient Real-Time Radial Distortion Correction for UAVs

Marcus Valtonen Ornhag, Patrik Persson, Marten Wadenback, Kalle Astrom, Anders Heyden; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1751-1760

In this paper we present a novel algorithm for onboard radial distortion correction for unmanned aerial vehicles (UAVs) equipped with an inertial measurement unit (IMU), that runs in real-time. This approach makes calibration procedures redundant, thus allowing for exchange of optics extemporaneously. By utilizing the IMU data, the cameras can be aligned with the gravity direction. This allows us to work with fewer degrees of freedom, and opens up for further intrinsic calibration. We propose a fast and robust minimal solver for simultaneously estimating the focal length, radial distortion profile and motion parameters from homographies. The proposed solver is tested on both synthetic and real data, and perform better or on par with state-of-the-art methods relying on pre-calibration procedures. Code available at: https://github.com/marcusvaltonen/HomLib.

****

Learned Dual-View Reflection Removal

Simon Niklaus, Xuaner (Cecilia) Zhang, Jonathan T. Barron, Neal Wadhwa, Rahul Garg, Feng Liu, Tianfan Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3713-3722

Traditional reflection removal algorithms either use a single image as input, which suffers from intrinsic ambiguities, or use multiple images from a moving camera, which is inconvenient for users. We instead propose a learning-based dereflection algorithm that uses stereo images as input. This is an effective trade-off between the two extremes: the parallax between two views provides cues to remove reflections, and two views are easy to capture due to the adoption of stereo cameras in smartphones. Our model consists of a learning-based reflection-invariant flow model for dual-view registration, and a learned synthesis model for combining aligned image pairs. Because no dataset for dual-view reflection removal exists, we render a synthetic dataset of dual-views with and without reflections for use in training. Our evaluation on an additional real-world dataset of stereo pairs shows that our algorithm outperforms existing single-image and multi-image dereflection approaches.

****

Goal-Driven Long-Term Trajectory Prediction

Hung Tran, Vuong Le, Truyen Tran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 796-805

The prediction of humans' short-term trajectories has advanced significantly with the use of powerful sequential modeling and rich environment feature extraction. However, long-term prediction is still a major challenge for the current methods as the errors could accumulate along the way. Indeed, consistent and stable prediction far to the end of a trajectory inherently requires deeper analysis into the overall structure of that trajectory, which is related to the pedestrian's intention on the destination of the journey. In this work, we propose to model a hypothetical process that determines pedestrians' goals and the impact of such process on long-term future trajectories. We design Goal-driven Trajectory Prediction model - a dual-channel neural network that realizes such intuition. The two channels of the network take their dedicated roles and collaborate to generate future trajectories. Different than conventional goal-conditioned, planning-based methods, the model architecture is designed to generalize the patterns and work across different scenes with arbitrary geometrical and semantic structures. The model is shown to outperform the state-of-the-art in various settings, especially in large prediction horizons. This result is another evidence for the effectiveness of adaptive structured representation of visual and geometrical features in human behavior analysis.

****

VideoSSL: Semi-Supervised Learning for Video Classification

Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, Hongcheng Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1110-1119

We propose a semi-supervised learning approach for video classification, VideoSSL, using convolutional neural networks (CNN). Like other computer vision tasks, existing supervised video classification methods demand a large amount of labeled data to attain good performance. However, annotation of a large dataset is expensive and time consuming. To minimize the dependence on a large annotated dataset, our proposed semi-supervised method trains from a small number of labeled examples and exploits two regulatory signals from unlabeled data. The first signal is the pseudo-labels of unlabeled examples computed from the confidences of the CNN being trained. The other is the normalized probabilities, as predicted by an image classifier CNN, that captures the information about appearances of the interesting objects in the video. We show that, under the supervision of these guiding signals from unlabeled examples, a video classification CNN can achieve impressive performances utilizing a small fraction of annotated examples on three publicly available datasets: UCF101, HMDB51 and Kinetics.

****

Mutual Information Maximization on Disentangled Representations for Differential Morph Detection

Sobhan Soleymani, Ali Dabouei, Fariborz Taherkhani, Jeremy Dawson, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1731-1741

In this paper, we present a novel differential morph detection framework, utilizing landmark and appearance disentanglement. In our framework, the face image is represented in the embedding domain using two disentangled but complementary representations. The network is trained by triplets of face images, in which the intermediate image inherits the landmarks from one image and the appearance from the other image. This initially trained network is further trained for each dataset using contrastive representations. We demonstrate that, by employing appearance and landmark disentanglement, the proposed framework can provide state-of-the-art differential morph detection performance. This functionality is achieved by the using distances in landmark, appearance, and ID domains. The performance of the proposed framework is evaluated using three morph datasets generated with different methodologies.

****

Few-Shot Learning via Feature Hallucination With Variational Inference

Qinxuan Luo, Lingfeng Wang, Jingguo Lv, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3963-3972

Deep learning has achieved huge success in the field of artificial intelligence, but the performance heavily depends on labeled data. Few-shot learning aims to make a model rapidly adapt to unseen classes with few labeled samples after training on a base dataset, and this is useful for tasks lacking labeled data such as medical image processing. Considering that the core problem of few-shot learning is the lack of samples, a straightforward solution to this issue is data augmentation. This paper proposes a generative model (VI-Net) based on a cosine-classifier baseline. Specifically, we construct a framework to learn to define a generating space for each category in the latent space based on few support samples. In this way, new feature vectors can be generated to help make the decision bo undary of classifier sharper during the fine-tuning process. To evaluate the effectiveness of our proposed approach, we perform comparative experiments and ablation studies on mini-ImageNet and CUB. Experimental results show that VI-Net does improve performance compared with the baseline and obtains the state-of-the-art result among other augmentation-based methods.

****

Line Art Correlation Matching Feature Transfer Network for Automatic Animation Colorization

Qian Zhang, Bo Wang, Wei Wen, Hai Li, Junhui Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3872-3881

Automatic animation line art colorization is a challenging computer vision problem, since the information of the line art is highly sparse and abstracted and there exists a strict requirement for the color and style consistency between frames. Recently, a lot of Generative Adversarial Network (GAN) based image-to-image translation methods for single line art colorization have emerged. They can generate perceptually appealing results conditioned on line art images. However, these methods can not be adopted for the purpose of animation colorization because there is a lack of consideration of the in-between frame consistency. Existing methods simply input the previous colored frame as a reference to color the next line art, which will mislead the colorization due to the spatial misalignment of the previous colored frame and the next line art especially at positions where apparent changes happen. To address these challenges, we design a kind of correlation matching feature transfer model (called CMFT) to align the colored reference feature in a learnable way and integrate the model into an U-Net based generator in a coarse-to-fine manner. This enables the generator to transfer the layer-wise synchronized features from the deep semantic code to the content progressively. Extension evaluation shows that CMFT model can effectively improve the in-between consistency and the quality of colored frames especially when the motion is intense and diverse.

****

Motion Adaptive Deblurring With Single-Photon Cameras

Trevor Seets, Atul Ingle, Martin Laurenzis, Andreas Velten; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1945-1954

Single-photon avalanche diodes (SPADs) are a rapidly developing image sensing technology with extreme low-light sensitivity and picosecond timing resolution. These unique capabilities have enabled SPADs to be used in applications like LiDAR, non-line-of-sight imaging and fluorescence microscopy that require imaging in photon-starved scenarios. In this work we harness these capabilities for dealing with motion blur in a passive imaging setting in low illumination conditions. Our key insight is that the data captured by a SPAD array camera can be represented as a 3D spatio-temporal tensor of photon detection events which can be integrated along arbitrary spatio-temporal trajectories with dynamically varying integration windows, depending on scene motion. We propose an algorithm that estimates pixel motion from photon timestamp data and dynamically adapts the integration windows to minimize motion blur. Our simulation results show the applicability of this algorithm to a variety of motion profiles including translation, rotation and local object motion. We also demonstrate the real-world feasibility of our method on data captured using a 32x32 SPAD camera.

****

Revisiting Batch Normalization for Improving Corruption Robustness

Philipp Benz, Chaoning Zhang, Adil Karjauv, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 494-503

The performance of DNNs trained on clean images has been shown to decrease when the test images have common corruptions. In this work, we interpret corruption robustness as a domain shift and propose to rectify batch normalization (BN) statistics for improving model robustness. This is motivated by perceiving the shift from the clean domain to the corruption domain as a style shift that is represented by the BN statistics. We find that simply estimating and adapting the BN st atistics on a few (32 for instance) representation samples, without retraining the model, improves the corruption robustness by a large margin on several benchmark datasets with a wide range of model architectures. For example, on ImageNet-C, statistics adaptation improves the top1 accuracy of ResNet50 from 39.2% to 48.7%. Moreover, we find that this technique can further improve state-of-the-art robust models from 58.1% to 63.3%.

****

Adaptive Streaming of 360-Degree Videos With Reinforcement Learning

Sohee Park, Minh Hoai, Arani Bhattacharya, Samir R. Das; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1839-1848

For bandwidth-efficient streaming of 360-degree videos, the streaming technique must adapt both to the changing viewport of the user and variations of the available network bandwidth. The state-of-the-art streaming techniques for this problem attempt to solve an optimization using simplified rules that do not adapt very well to the uncertainties related to the viewport or network. We adopt a 3D-Convolutional Neural Networks (3DCNN) model to extract spatio-temporal features of videos and predict the viewport. Given the sequential decision-making nature of such streaming technique, we then apply a Reinforcement Learning (RL) based adaptive streaming approach. We address the challenges of using RL in this scenario, such as large action space and delayed reward evaluation. Comprehensive evaluations with real network traces show that the proposed method outperforms three tile-based streaming techniques for 360-degree videos. Compared to the tile-based streaming techniques, the average user-perceived bitrate of the proposed method is 1.3-1.7 times higher and the average quality of experience of the proposed method is also 1.6-3.4 times higher. Subjective user studies further confirm the superiority of the proposed approach.

****

Automatic Calibration of the Fisheye Camera for Egocentric 3D Human Pose Estimation From a Single Image

Yahui Zhang, Shaodi You, Theo Gevers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1772-1781

We propose a method for egocentric 3D human pose estimation from a single image captured by a fisheye camera. The problem of estimating the egocentric 3D pose for a fisheye camera is that images may be subject to strong image distortions (e.g. 2D poses on the image plane that pass through the line of sight of the fisheye lens). Therefore, in this paper, we approach this problem by an automatic calibration module. Given a single image, our network first estimates 3D joint locations of a human in camera coordinates. To alleviate the impact of image distortions on 3D human pose estimation, we then use the automatic calibration to further regularize the 3D predictions. Experimental results demonstrate that the proposed method achieves state-of-the-art performance.

****

Two-Level Adversarial Visual-Semantic Coupling for Generalized Zero-Shot Learning

Shivam Chandhok, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3100-3108

The performance of generative zero-shot methods mainly depends on the quality of generated features and how well the model facilitates knowledge transfer between visual and semantic domains. The quality of generated features is a direct consequence of the ability of the model to capture the several modes of the underlying data distribution. To address these issues, we propose a new two-level joint maximization idea to augment the generative network with an inference network during training which helps our model capture the several modes of the data and generate features that better represent the underlying data distribution. This provides strong cross-modal interaction for effective transfer of knowledge between visual and semantic domains. Furthermore, existing methods train the zero-shot classifier either on generate synthetic image features or latent embeddings produced by leveraging representation learning. In this work, we unify these paradigms into a single model which in addition to synthesizing image features, also u tilizes the representation learning capabilities of the inference network to provide discriminative features for the final zero-shot recognition task. We evaluate our approach on four benchmark datasets i.e. CUB, FLO, AWA1 and AWA2 against several state-of-the-art methods, and show its performance. We also perform ablation studies to analyze and understand our method more carefully for the Generalized Zero-shot Learning task.

****

Unsupervised Domain Adaptation in Semantic Segmentation via Orthogonal and Clustered Embeddings

Marco Toldo, Umberto Michieli, Pietro Zanuttigh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1358-1368

Deep learning frameworks allowed for a remarkable advancement in semantic segmentation, but the data hungry nature of convolutional networks has rapidly raised the demand for adaptation techniques able to transfer learned knowledge from label-abundant domains to unlabeled ones. In this paper we propose an effective Unsupervised Domain Adaptation (UDA) strategy, based on a feature clustering method that captures the different semantic modes of the feature distribution and groups features of the same class into tight and well-separated clusters. Furthermore, we introduce two novel learning objectives to enhance the discriminative clustering performance: an orthogonality loss forces spaced out individual representations to be orthogonal, while a sparsity loss reduces class-wise the number of active feature channels. The joint effect of these modules is to regularize the structure of the feature space. Extensive evaluations in the synthetic-to-real scenario show that we achieve state-of-the-art performance.

****

Saliency Driven Perceptual Image Compression

Yash Patel, Srikar Appalaraju, R. Manmatha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 227-236

This paper proposes a new end-to-end trainable model for lossy image compression, which includes several novel components. The method incorporates 1) an adequate perceptual similarity metric; 2) saliency in the images; 3) a hierarchical auto-regressive model. This paper demonstrates that the popularly used evaluations metrics such as MS-SSIM and PSNR are inadequate for judging the performance of image compression techniques as they do not align with the human perception of similarity. Alternatively, a new metric is proposed, which is learned on perceptual similarity data specific to image compression. The proposed compression model incorporates the salient regions and optimizes on the proposed perceptual similarity metric. The model not only generates images which are visually better but also gives superior performance for subsequent computer vision tasks such as object detection and segmentation when compared to existing engineered or learned compression techniques.

****

A Multi-Task Learning Approach for Human Activity Segmentation and Ergonomics Risk Assessment

Behnoosh Parsa, Ashis G. Banerjee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2352-2362

We propose a new approach to Human Activity Evaluation (HAE) in long videos using graph-based multi-task modeling. Previous works in activity evaluation either directly compute a metric using a detected skeleton or use the scene information to regress the activity score. These approaches are insufficient for accurate activity assessment since they only compute an average score over a clip, and do not consider the correlation between the joints and body dynamics. Moreover, they are highly scene-dependent which makes the generalizability of these methods questionable. We propose a novel multi-task framework for HAE that utilizes a Graph Convolutional Network backbone to embed the interconnections between human joints in the features. In this framework, we solve the Human Activity Segmentation (HAS) problem as an auxiliary task to improve activity assessment. The HAS head is powered by an encoder-Decoder Temporal Convolutional Network to semantically segment long videos into distinct activity classes, whereas, HAE uses a Long-Short-Term-Memory-based architecture. We evaluate our method on the UW-IOM and TU

M Kitchen datasets and discuss the success and failure cases in these two datasets.

****

A Variational Information Bottleneck Based Method to Compress Sequential Networks for Human Action Recognition

Ayush Srivastava, Oshin Dutta, Jigyasa Gupta, Sumeet Agarwal, Prathosh AP; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2745-2754

In the last few years, deep neural networks' compression has become an important strand of machine learning and computer vision research. Deep models require sizeable computational complexity and storage when used, for instance, for Human Action Recognition (HAR) from videos, making them unsuitable to be deployed on edge devices. In this paper, we address this issue and propose a method to effectively compress Recurrent Neural Networks (RNNs) such as Gated Recurrent Units (GRUs) and Long-Short-Term-Memory Units (LSTMs) that are used for HAR. We use a Variational Information Bottleneck (VIB) theory-based pruning approach to limit the information flow through the sequential cells of RNNs to a small subset. Further, we combine our pruning method with a specific group-lasso regularization technique that significantly improves compression. The proposed techniques reduce model parameters and memory footprint from latent representations, with little or no reduction in the validation accuracy while increasing the inference speed several-fold. We perform experiments on the three widely used Action Recognition datasets, viz. UCF11, HMDB51, and UCF101, to validate our approach. We show that our method achieves over 70 times greater compression than the nearest competitor with comparable accuracy for action recognition on UCF11.

****

Adversarial Reinforcement Learning for Unsupervised Domain Adaptation

Youshan Zhang, Hui Ye, Brian D. Davison; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 635-644

Transferring knowledge from an existing labeled domain to a new domain often suffers from domain shift in which performance degrades because of differences between the domains. Domain adaptation has been a prominent method to mitigate such a problem. There have been many pre-trained neural networks for feature extraction. However, little work discusses how to select the best feature instances across different pre-trained models for both the source and target domain. We propose a novel approach to select features by employing reinforcement learning, which learns to select the most relevant features across two domains. Specifically, in this framework, we employ Q-learning to learn policies for an agent to make feature selection decisions by approximating the action-value function. After selecting the best features, we propose an adversarial distribution alignment learning to improve the prediction results. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art methods.

****

A Robust and Efficient Framework for Sports-Field Registration

Xiaohan Nie, Shixing Chen, Raffay Hamid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1936-1944

We propose a novel framework to register sports-fields as they appear in broadcast sports videos. Unlike previous approaches, we particularly address the challenge of field-registration when: (a) there are not enough distinguishable features on the field, and (b) no prior knowledge is available about the camera. To this end, we detect a grid of keypoints distributed uniformly on the entire field instead of using only sparse local corners and line intersections, thereby extending the keypoint coverage to the texture-less parts of the field as well. To further improve keypoint based homography estimate, we differentialbly warp and align it with a set of dense field-features defined as normalized distance-map of pixels to their nearest lines and key-regions. We predict the keypoints and dense field-features simultaneously using a multi-task deep network to achieve computational efficiency. To have a comprehensive evaluation, we have compiled a new dataset called SportsFields which is collected from 192 video-clips from 5 different sports covering large environmental and camera variations. We empirically de monstrate that our algorithm not only achieves state of the art field-registration accuracy but also runs in real-time for HD resolution videos using commodity hardware.

****

Robust Lensless Image Reconstruction via PSF Estimation

Joshua D. Rego, Karthik Kulkarni, Suren Jayasuriya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 403-412

Lensless imaging is a new, emerging modality where image sensors utilize optical elements in front of the sensor to perform multiplexed imaging. There have been several recent papers to reconstruct images from lensless imagers, including methods that utilize deep learning for state-of-the-art performance. However, many of these methods require explicit knowledge of the optical element, such as the point spread function, or learn the reconstruction mapping for a single fixed PSF. In this paper, we explore a neural network architecture that performs joint image reconstruction and PSF estimation to robustly recover images captured with multiple PSFs from different cameras. Using adversarial learning, this approach achieves improved reconstruction results that do not require explicit knowledge of the PSF at test-time and shows an added improvement in the reconstruction model's ability to generalize to variations in the camera's PSF. This allows lensless cameras to be utilized in a wider range of applications that require multiple cameras without the need to explicitly train a separate model for each new camera.

****

Person-in-Context Synthesis With Compositional Structural Space

Weidong Yin, Ziwei Liu, Leonid Sigal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2827-2836

Despite significant progress, controlled generation of complex images with interacting people remains difficult. Existing layout generation methods fall short of synthesizing realistic person instances; while pose-guided generation approaches focus on a single person and assume simple or known backgrounds. To tackle these limitations, we propose a new problem, person in context synthesis, which aims to synthesize diverse person instance(s) in consistent contexts, with user control over both. The context is specified by the bounding box object layout which lacks shape information, while pose of the person(s) by keypoints which are sparsely annotated. To handle the stark difference in input structures, we proposed two separate neural branches to attentively composite the respective (context/person) inputs into shared "compositional structural space", which encodes shape, location and appearance information for both context and person structures in a disentangled manner. This structural space is then decoded to the image space using multi-level feature modulation strategy, and learned in a self supervised manner from image collections and their corresponding inputs. Extensive experiments on two large-scale datasets (COCO-Stuff and Visual Genome) demonstrate that our framework outperforms state-of-the-art methods w.r.t. synthesis quality.

****

Deep Template-Based Object Instance Detection

Jean-Philippe Mercier, Mathieu Garon, Philippe Giguere, Jean-Francois Lalonde; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1507-1516

Much of the focus in the object detection literature has been on the problem of identifying the bounding box of a particular class of object in an image. Yet, in contexts such as robotics and augmented reality, it is often necessary to find a specific object instance--a unique toy or a custom industrial part for example--rather than a generic object class. Here, applications can require a rapid shift from one object instance to another, thus requiring fast turnaround which affords little-to-no training time. What is more, gathering a dataset and training a model for every new object instance to be detected can be an expensive and time-consuming process. In this context, we propose a generic 2D object instance detection approach that uses example viewpoints of the target object at test time to retrieve its 2D location in RGB images, without requiring any additional training (i.e. fine-tuning) step. To this end, we present an end-to-end architectur

e that extracts global and local information of the object from its viewpoints. The global information is used to tune early filters in the backbone while local viewpoints are correlated with the input image. Our method offers an improvement of almost 30 mAP over the previous template matching methods on the challenging Occluded Linemod [3] dataset (overall mAP of 50.7). Our experiments also show that our single generic model (not trained on any of the test objects) yields detection results that are on par with approaches that are trained specifically on the target objects.
****

Future Moment Assessment for Action Query
Qiuhong Ke, Mario Fritz, Bernt Schiele; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3219-3228
In this paper, we aim to tackle the task of Assessing Future Moment of an Action of Interest (AFM-AI). The goal of this task is to assess if an action of interest will happen or not as well as the starting moment of the action. We aim to assess starting moments at any time-horizon of the future. To this end, we tackle the regression task of the starting moments as a generation task using a Deterministic Residual Guided Variational Regression Module (DR-VRM), which is built on a Variational Regression Module (VRM) and a deterministic residual network. The VRM takes the uncertainty into account and is capable of generating diverse predictions for the starting moment. The deterministic network encourages the VRM to learn from deterministic residual information in order to generate more precise predictions for moment assessment. Experimental results on three datasets clearly show that the proposed method is capable of generating both diverse and precise predictions of starting moments for query actions.
****

Fast Fourier Intrinsic Network
Yanlin Qian, Miaojing Shi, Joni-Kristian Kamarainen, Jiri Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3169-3178
We address the problem of decomposing an image into albedo and shading. We propose the Fast Fourier Intrinsic Network, FFI-Net in short, that operates in the spectral domain, splitting the input into several spectral bands. Weights in FFI-Net are optimized in the spectral domain, allowing faster convergence to a lower error. FFI-Net is lightweight and does not need auxiliary networks for training. The network is trained end-to-end with a novel spectral loss which measures the global distance between the network prediction and corresponding ground truth. FFI-Net achieves state-of-the-art performance on MPI-Sintel, MIT Intrinsic, and IIW datasets.
****

Pretraining Boosts Out-of-Domain Robustness for Pose Estimation
Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, Mackenzie W. Mathis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1859-1868
Neural networks are highly effective tools for pose estimation. However, as in other computer vision tasks, robustness to out-of-domain data remains a challenge, especially for small training sets that are common for real-world applications. Here, we probe the generalization ability with three architecture classes (MobileNetV2s, ResNets, and EfficientNets) for pose estimation. We developed a dataset of 30 horses that allowed for both "within-domain" and "out-of-domain" (unseen horse) benchmarking---this is a crucial test for robustness that current human pose estimation benchmarks do not directly address. We show that better ImageNet-performing architectures perform better on both within- and out-of-domain data if they are first pretrained on ImageNet. We additionally show that better ImageNet models generalize better across animal species. Furthermore, we introduce Horse-C, a new benchmark for common corruptions for pose estimation, and confirm that pretraining increases performance in this domain shift context as well. Overall, our results demonstrate that transfer learning is beneficial for out-of-domain robustness.
****

We Don't Need Thousand Proposals: Single Shot Actor-Action Detection in Videos

Aayush J. Rana, Yogesh S. Rawat; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2960-2969

We propose SSA2D, a simple yet effective end-to-end deep network for actor-action detection in videos. The existing methods take a top-down approach based on region-proposals (RPN), where the action is estimated based on the detected proposals followed by post-processing such as non-maximal suppression. While effective in terms of performance, these methods pose limitations in scalability for dense video scenes with a high memory requirement for thousands of proposals. We propose to solve this problem from a different perspective where we don't need any proposals. SSA2D is a unified network, which performs pixel level joint actor-action detection in a single-shot, where every pixel of the detected actor is assigned an action label. SSA2D has two main advantages: 1) It is a fully convolutional network which does not require any proposals and post-processing making it memory as well as time efficient, 2) It is easily scalable to dense video scenes as its memory requirement is independent of the number of actors present in the scene. We evaluate the proposed method on the Actor-Action dataset (A2D) and Video Object Relation (VidOR) dataset, demonstrating its effectiveness in multiple actors and action detection in a video. SSA2D is 11x faster during inference with comparable (sometimes better) performance and fewer network parameters when compared with the prior works. Code available at https://github.com/aayushjr/ssa2d

****

AutoRetouch: Automatic Professional Face Retouching

Alireza Shafaei, James J. Little, Mark Schmidt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 990-998

Face retouching is one of the most time-consuming steps in professional photography pipelines. The existing automated approaches blindly apply smoothing on the skin, destroying the delicate texture of the face. We present the first automatic face retouching approach that produces high-quality professional-grade results in less than two seconds. Unlike previous work, we show that our method preserves textures and distinctive features while retouching the skin. We demonstrate that our trained models generalize across datasets and are suitable for low-resolution cellphone images. Finally, we release the first large-scale, professionally retouched dataset with our baseline to encourage further work on the presented problem.

****

Hierarchical Generative Adversarial Networks for Single Image Super-Resolution

Weimin Chen, Yuqing Ma, Xianglong Liu, Yi Yuan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 355-364

Recently, deep convolutional neural network (CNN) have achieved promising performance for single image super-resolution (SISR). However, they usually extract features on a single scale and lack sufficient supervision information, leading to undesired artifacts and unpleasant noise in super-resolution (SR) images. To address this problem, we first propose a hierarchical feature extraction module (HFEM) to extract the features in multiple scales, which helps concentrate on both local textures and global semantics. Then, a hierarchical guided reconstruction module (HGRM) is introduced to reconstruct more natural structural textures in SR images via intermediate supervisions in a progressive manner. Finally, we integrate HFEM and HGRM in a simple yet efficient end-to-end framework named hierarchical generative adversarial networks (HSRGAN) to recover consistent details, and thus obtain the semantically reasonable and visually realistic results. Extensive experiments on five common datasets demonstrate that our method shows favorable visual quality and superior quantitative performance compared to state-of-the-art methods for SISR.

****

Text-to-Image Generation Grounded by Fine-Grained User Attention

Jing Yu Koh, Jason Baldridge, Honglak Lee, Yinfei Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 237-246

Localized Narratives is a dataset with detailed natural language descriptions of

images paired with mouse traces that provide a sparse, fine-grained visual grounding for phrases. We propose TReCS, a sequential model that exploits this grounding to generate images. TReCS uses descriptions to retrieve segmentation masks and predict object labels aligned with mouse traces. These alignments are used to select and position masks to generate a fully covered segmentation canvas; the final image is produced by a segmentation-to-image generator using this canvas. This multi-step, retrieval-based approach outperforms existing direct text-to-image generation models on both automatic metrics and human evaluations: overall, its generated images are more photo-realistic and better match descriptions.
****

maskedFaceNet: A Progressive Semi-Supervised Masked Face Detector
Shitala Prasad, Yiqun Li, Dongyun Lin, Dong Sheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3389-3398
To reduce the risk of infecting or being infected by the recent COVID-19 virus, wearing mask is enforced or recommended by many countries. AI based system for automatically detecting whether individuals are wearing face mask becomes an urgent requirement in high risk facilities and crowded public places. Due to lacking of existing masked face datasets and the urgent low-cost application requirement, we propose a progressive semi-supervised learning method - called maskedFaceNet to minimize the efforts on data annotation and letting deep models to learn by using less annotated training data. With this method, the detection accuracy is further improved progressively while adapting to various application scenarios. Experimental results show that our maskedFaceNet is more efficient and accurate compared to other methods. Furthermore, we also contribute two masked face datasets for benchmarking and for the benefit of future research.
****

Adaptive Privacy Preserving Deep Learning Algorithms for Medical Data
Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen T.C. Wong, Hien Van Nguyen, Miao Pan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1169-1178
Deep learning holds a great promise of revolutionizing healthcare and medicine. Unfortunately, various inference attack models demonstrated that deep learning puts sensitive patient information at risk. The high capacity of deep neural networks is the main reason behind the privacy loss. In particular, patient information in the training data can be unintentionally memorized by a deep network. Adversarial parties can extract that information given the ability to access or query the network. In this paper, we propose a novel privacy-preserving mechanism for training deep neural networks. Our approach adds decaying Gaussian noise to the gradients at every training iteration. This is in contrast to the mainstream approach adopted by Google's TensorFlow Privacy, which employs the same noise scale in each step of the whole training process. Compared to existing methods, our proposed approach provides an explicit closed-form mathematical expression to approximately estimate the privacy loss. It is easy to compute and can be useful when the users would like to decide proper training time, noise scale, and sampling ratio during the planning phase. We provide extensive experimental results using one real-world medical dataset (chest radiographs from the CheXpert dataset) to validate the effectiveness of the proposed approach. The proposed differential privacy based deep learning model achieves significantly higher classification accuracy over the existing methods with the same privacy budget.
****

Automatic Object Recoloring Using Adversarial Learning
Siavash Khodadadeh, Saeid Motiian, Zhe Lin, Ladislau Boloni, Shabnam Ghadar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1488-1496
We propose a novel method for automatic object recoloring based on Generative Adversarial Networks (GANs). The user can simply give commands of the form ""recolor <object> to <color>"" which will be executed without any need of manual edit. Our approach takes advantage of pre-trained object detectors and saliency mask segmentation networks. The segmented mask of the given object along with the target color and the original image form the input to the GAN. The use of cycle con

sistency loss ensures the realistic look of the results. To our best knowledge, this is the first algorithm where the automatic recoloring is only limited by the ability of the mask extractor to map a natural language tag to a specific object in the image (several hundred object types at the time of this writing). For a performance comparison, we also adapted other state of the art methods to perform this task. We found that our method had consistently yielded qualitatively better recoloring results.

****

Improved Training of Generative Adversarial Networks Using Decision Forests

Yan Zuo, Gil Avraham, Tom Drummond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3492-3501

Whilst Generative Adversarial Networks (GANs) have gained a reputation as powerful generative models, they are notoriously difficult to train and suffer from instability in optimisation. Recent methods for tackling this drawback have typically approached it by inducing better behaviour on the discriminator component of the GAN; these include loss function modification, gradient regularisation and weight normalisation to create a discriminator that is well-behaved from a Lipschitz perspective. In this paper, we propose a novel and orthogonal contribution which modifies the architecture of a GAN. Our method embeds the powerful discriminating capabilities inherent in decision forests within the discriminator of a GAN. Empirically, we test the effectiveness of our approach on the CIFAR-10, Oxford Flowers and CUB Birds datasets. We show that our technique is easy to incorporate into existing GAN baselines and offers improvements on Frechet-Inception Distance (FID) scores by as high as 56.1% over several GAN baselines.

****

Boosting Monocular Depth With Panoptic Segmentation Maps

Faraz Saeedan, Stefan Roth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3853-3862

Monocular depth prediction is ill-posed by nature; hence successful approaches need to exploit the available cues to the fullest. Yet, real-world training data with depth ground-truth suffers from limited variability, and data acquired from depth sensors is also sparse and prone to noise. While available datasets with semantic annotations might help to better exploit semantic cues, they are not immediately usable for depth prediction. We show how to leverage panoptic segmentation maps to boost monocular depth predictors in stereo training setups. In particular, we augment a self-supervised training scheme through panoptic-guided smoothing, panoptic-guided alignment, and panoptic left-right consistency from ground truth or inferred panoptic segmentation maps. Our approach incurs only a minor overhead, can easily be applied to a wide range of depth estimation methods that are trained at least partially using stereo pairs, providing a substantial boost in accuracy.

****

Context-Aware Domain Adaptation in Semantic Segmentation

Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, Junzhou Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 514-524

In this paper, we consider the problem of unsupervised domain adaptation in the semantic segmentation. There are two primary issues in this field, i.e., what and how to transfer domain knowledge across two domains. Existing methods mainly focus on adapting domain-invariant features (what to transfer) through adversarial learning (how to transfer). Context dependency is essential for semantic segmentation, however, its transferability is still not well understood. Furthermore, how to transfer contextual information across two domains remains unexplored. Motivated by this, we propose a cross-attention mechanism based on self-attention to capture context dependencies between two domains and adapt transferable context. To achieve this goal, we design two cross-domain attention modules to adapt context dependencies from both spatial and channel views. Specifically, the spatial attention module captures local feature dependencies between each position in the source and target image. The channel attention module models semantic dependencies between each pair of cross-domain channel maps. To adapt context depen dencies, we further selectively aggregate the context information from two domains. The superiority of our method over existing state-of-the-art methods is empirically proved on "GTA5 to Cityscapes" and "SYNTHIA to Cityscapes".

****

Neural Contrast Enhancement of CT Image

Minkyo Seo, Dongkeun Kim, Kyungmoon Lee, Seunghoon Hong, Jae Seok Bae, Jung Hoon Kim, Suha Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3973-3982

Contrast materials are often injected into body to contrast specific tissues in Computed Tomography (CT) images. Contrast Enhanced CT (CECT) images obtained in this way are more useful than Non-Enhanced CT (NECT) images for medical diagnosis, but not available for everyone due to side effects of the contrast materials. Motivated by this, we develop a neural network that takes NECT images and generates their CECT counterparts. Learning such a network is extremely challenging since NECT and CECT images for training are not aligned even at the same location of the same patient due to movements of internal organs. We propose a two-stage framework to address this issue. The first stage trains an auxiliary network that removes the effect of contrast enhancement in CECT images to synthesize their NECT counterparts well-aligned with them. In the second stage, the target model is trained to predict the real CECTimages given a synthetic NECT image as input. Experimental results and analysis by physicians on abdomen CT images suggest that our method outperforms existing models for neural image synthesis.

****

TB-Net: A Three-Stream Boundary-Aware Network for Fine-Grained Pavement Disease Segmentation

Yujia Zhang, Qianzhong Li, Xiaoguang Zhao, Min Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3655-3664

Regular pavement inspection plays a significant role in road maintenance for safety assurance. Existing methods mainly address the tasks of crack detection and segmentation that are only tailored for long-thin crack disease. However, there are many other types of diseases with a wider variety of sizes and patterns that are also essential to segment in practice, bringing more challenges towards fine-grained pavement inspection. In this paper, our goal is not only to automatically segment cracks, but also to segment other complex pavement diseases as well as typical landmarks (markings, runway lights, etc.) and commonly seen water/oil stains in a single model. To this end, we propose a three-stream boundary-aware network (TB-Net). It consists of three streams fusing the low-level spatial and the high-level contextual representations as well as the detailed boundary information. Specifically, the spatial stream captures rich spatial features. The context stream, where an attention mechanism is utilized, models the contextual relationships over local features. The boundary stream learns detailed boundaries using a global-gated convolution to further refine the segmentation outputs. The network is trained using a dual-task loss in an end-to-end manner, and experiments on a newly collected fine-grained pavement disease dataset show the effectiveness of our TB-Net.

****

Alleviating Over-Segmentation Errors by Detecting Action Boundaries

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, Hirokatsu Kataoka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2322-2331

We propose an effective framework for the temporal action segmentation task, namely an Action Segment Refinement Framework (ASRF). Our model architecture consists of a long-term feature extractor and two branches: the Action Segmentation Branch (ASB) and the Boundary Regression Branch (BRB). The long-term feature extractor provides shared features for the two branches with a wide temporal receptive field. The ASB classifies video frames with action classes, while the BRB regresses the action boundary probabilities. The action boundaries predicted by the BRB refine the output from the ASB, which results in a significant performance improvement. Our contributions are three-fold: (i) We propose a framework for temporal action segmentation, the ASRF, which divides temporal action segmentation into frame-wise action classification and action boundary regression. Our framework refines frame-level hypotheses of action classes using predicted action boundaries. (ii) We propose a loss function for smoothing the transition of action probabilities, and analyze combinations of various loss functions for temporal action segmentation. (iii) Our model outperforms state-of-the-art methods on three challenging datasets, offering an improvement of up to 13.7% in terms of segmental edit distance and up to 16.1% in terms of segmental F1 score. Our code is publicly available.

****

Towards Resolving the Challenge of Long-Tail Distribution in UAV Images for Object Detection

Weiping Yu, Taojiannan Yang, Chen Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3258-3267

Existing methods for object detection in UAV images ignored an important challenge -- imbalanced class distribution -- which leads to poor performance on tail classes. We systematically investigate existing solutions to long-tail problems and unveil that re-balancing methods that are effective on natural image datasets cannot be trivially applied to UAV datasets. To this end, we rethink long-tailed object detection in UAV images and propose the Dual Sampler and Head detection Network (DSHNet), which is the first work that aims to resolve long-tail distribution in UAV images. The key components in DSHNet include Class-Biased Samplers (CBS) and Bilateral Box Heads (BBH), which are developed to cope with tail classes and head classes in a dual-path manner. Without bells and whistles, DSHNet significantly boosts the performance of tail classes on different detection frameworks. Moreover, DSHNet significantly outperforms base detectors and generic approaches for long-tail problems on VisDrone and UAVDT datasets. It achieves a new state-of-the-art performance when combining with image cropping methods.

****

Making DensePose Fast and Light

Ruslan Rakhimov, Emil Bogomolov, Alexandr Notchenko, Fung Mao, Alexey Artemov, Denis Zorin, Evgeny Burnaev; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1869-1877

DensePose estimation task is a significant step forward for enhancing user experience computer vision applications ranging from augmented reality to cloth fitting. Existing neural network models capable of solving this task are heavily parameterized and a long way from being transferred to an embedded or mobile device. To enable Dense Pose inference on the end device with current models, one needs to support an expensive server-side infrastructure and have a stable internet connection. To make things worse, mobile and embedded devices do not always have a powerful GPU inside. In this work, we target the problem of redesigning the DensePose R-CNN model's architecture so that the final network retains most of its accuracy but becomes more light-weight and fast. To achieve that, we tested and incorporated many deep learning innovations from recent years, specifically performing an ablation study on 23 efficient backbone architectures, multiple two-stage detection pipeline modifications, and custom model quantization methods. As a result, we achieved 17 times model size reduction and 2 times latency improvement compared to the baseline model.

****

MVHM: A Large-Scale Multi-View Hand Mesh Benchmark for Accurate 3D Hand Pose Estimation

Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 836-845

Estimating 3D hand poses from a single RGB image is challenging because depth ambiguity leads the problem ill-posed. Training hand pose estimators with 3D hand mesh annotations and multi-view images often results in significant performance gains. However, existing multi-view datasets are relatively small with hand joints annotated by off-the-shelf trackers or automated through model predictions, both which may be inaccurate and can introduce biases. Collecting a large-scale multi-view 3D hand pose images with accurate mesh and joint annotations is valuab le but strenuous. In this paper, we design a spin match algorithm that enables rigid mesh model matching without any target mesh ground truth. Based on the match algorithm, we propose an efficient pipeline to generate a large-scale multi-view hand mesh (MVHM) dataset with accurate 3D hand mesh and joint labels. We further present a multi-view hand pose estimation approach to verify that training a hand pose estimator with our generated dataset greatly enhances the performance. Experimental results show that our approach achieves the performance of 0.990 in \text AUC _ \text 20-50   on the MHP dataset compared to the previous state-of-the-art of 0.939 on this dataset.

****

Fast Pose Graph Optimization via Krylov-Schur and Cholesky Factorization

Gabriel Moreira, Manuel Marques, Joao Paulo Costeira; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1898-1906

Pose Graph Optimization (PGO) is an important problem in Computer Vision, particularly in motion estimation, whose objective consists of finding the rigid transformations that achieve the best global alignment of visual data on a common reference frame. The vast majority of PGO approaches rely on iterative techniques which refine an initial estimate until convergence is achieved. On the other hand, recent works have identified a global constraint which has cast this problem into the matrix completion domain. The success which both these formulations have had in computing accurate solutions efficiently has been overshadowed by large-scale industrial applications such as autonomous flight, self-driving cars and smart-cities, where it is necessary to fuse numerous images covering large areas but where each one of them has few pairwise observations. We propose a highly efficient algorithm to solve PGO which leverages the sparsity of the data by combining the Krylov-Schur method for spectral decomposition with Cholesky LDL factorization. Our method allows for high scalability, low computational cost and high precision, simultaneously.

****

Dense 3D-Reconstruction From Monocular Image Sequences for Computationally Constrained UAS

Matthias Domnik, Pedro Proenca, Jeff Delaune, Jorg Thiem, Roland Brockers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1820-1828

The ability to find safe landing sites over complex 3D terrain is an essential safety feature for fully autonomous small unmanned aerial systems (UAS), which requires on-board perception for 3D reconstruction and terrain analysis if the overflown terrain is unknown. This is a challenge for UAS that are limited in size, weight and computational power, such as small rotorcrafts executing autonomous missions on Earth, or in planetary applications such as the Mars Helicopter. For such a computationally constraint system, we propose a structure from motion approach that uses inputs from a single downward facing camera to produce dense point clouds of the overflown terrain in real time. In contrast to existing approaches, our method uses metric pose information from a visual-inertial odometry algorithm as camera pose priors, which allows deploying a fast pose refinement step to align camera frames such that a conventional stereo algorithm can be used for dense 3D reconstruction. We validate the performance of our approach with extensive evaluations in simulation, and demonstrate the feasibility with data from UAS flights.

****

DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video

Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1079-1088

This paper studies the task of temporal moment localization in a long untrimmed video using natural language query. Given a query sentence, the goal is to determine the start and end of the relevant segment within the video. Our key innovation is to learn a video feature embedding through a language-conditioned message

-passing algorithm suitable for temporal moment localization which captures the relationships between humans, objects and activities in the video. These relationships are obtained by a spatial subgraph that contextualized the scene representation using detected objects and human features. Moreover, a temporal sub-graph captures the activities within the video through time. Our method is evaluated on three standard benchmark datasets, and we also introduce YouCookII as a new benchmark for this task. Experiments show our method outperforms state-of-the-art methods on these datasets, confirming the effectiveness of our approach

****

ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning

Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, Lennart Svensson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1369-1378

The state of the art in semantic segmentation is steadily increasing in performance, resulting in more precise and reliable segmentations in many different applications. However, progress is limited by the cost of generating labels for training, which sometimes requires hours of manual labor for a single image. Because of this, semi-supervised methods have been applied to this task, with varying degrees of success. A key challenge is that common augmentations used in semi-supervised classification are less effective for semantic segmentation. We propose a novel data augmentation mechanism called ClassMix, which generates augmentations by mixing unlabelled samples, by leveraging on the network's predictions for respecting object boundaries. We evaluate this augmentation technique on two common semi-supervised semantic segmentation benchmarks, showing that it attains state-of-the-art results. Lastly, we also provide extensive ablation studies comparing different design decisions and training regimes.

****

Deep Image Compositing

He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 365-374

Image compositing is a task of combining regions from different images to compose a new image. A common use case is background replacement of portrait images. To obtain high quality composites, professionals typically manually perform multiple editing steps such as segmentation, matting and foreground color decontamination, which is very time consuming even with sophisticated photo editing tools. In this paper, we propose a new method which can automatically generate high-quality image compositing without any user input. Our method can be trained end-to-end to optimize exploitation of contextual and color information of both foreground and background images, where the compositing quality is considered in the optimization. Specifically, inspired by Laplacian pyramid blending, a denseconnected multi-stream fusion network is proposed to effectively fuse the information from the foreground and background images at different scales. In addition, we introduce a self-taught strategy to progressively train from easy to complex cases to mitigate the lack of training data. Experiments show that the proposed method can automatically generate high-quality composites and outperforms existing methods both qualitatively and quantitatively.

****

Novel View Synthesis via Depth-Guided Skip Connections

Yuxin Hou, Arno Solin, Juho Kannala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3119-3128

We introduce a principled approach for synthesizing new views of a scene given a single source image. Previous methods for novel view synthesis can be divided into image-based rendering methods (e.g, flow prediction) or pixel generation methods. Flow predictions enable the target view to re-use pixels directly, but can easily lead to distorted results. Directly regressing pixels can produce structurally consistent results but generally suffer from the lack of low-level details. In this paper, we utilize an encoder-decoder architecture to regress pixels of a target view. In order to maintain details, we couple the decoder aligned feature maps with skip connections, where the alignment is guided by predicted dept h map of the target view. Our experimental results show that our method does not suffer from distortions and successfully preserves texture details with aligned skip connections.

****

GlocalNet: Class-Aware Long-Term Human Motion Synthesis

Neeraj Battan, Yudhik Agrawal, Sai Soorya Rao, Aman Goel, Avinash Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 879-888

Synthesis of long-term human motion skeleton sequences is essential to aid human-centric video generation with potential applications in Augmented Reality, 3D character animations, pedestrian trajectory prediction, etc. Long-term human motion synthesis is a challenging task due to multiple factors like, long-term temporal dependencies among poses, cyclic repetition across poses, bi-directional and multi-scale dependencies among poses, variable speed of actions, and a large as well as partially overlapping space of temporal pose variations across multiple class/types of human activities. This paper aims to address these challenges to synthesize a long-term (>6000 ms) human motion trajectory across a large variety of human activity classes (>50). We propose a two-stage activity generation method to achieve this goal, where the first stage deals with learning the long-term global pose dependencies in activity sequences by learning to synthesize a sparse motion trajectory while the second stage addresses the generation of dense motion trajectories taking the output of the first stage. We demonstrate the superiority of the proposed method over SOTA methods using various quantitative evaluation metrics on publicly available datasets.

****

JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition

Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, Jiangbo Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2735-2744

Skeleton-based action recognition has attracted research attentions in recent years. One common drawback in currently popular skeleton-based human action recognition methods is that the sparse skeleton information alone is not sufficient to fully characterize human motion. This limitation makes several existing methods incapable of correctly classifying action categories which exhibit only subtle motion differences. In this paper, we propose a novel framework for employing human pose skeleton and joint-centered light-weight information jointly in a two-stream graph convolutional network, namely, JOLO-GCN. Specifically, we use Joint-aligned optical Flow Patches (JFP) to capture the local subtle motion around each joint as the pivotal joint-centered visual information. Compared to the pure skeleton-based baseline, this hybrid scheme effectively boosts performance, while keeping the computational and memory overheads low. Experiments on the NTU RGB+D, NTU RGB+D 120, and the Kinetics-Skeleton dataset demonstrate clear accuracy improvements attained by the proposed method over the state-of-the-art skeleton-based methods.

****

Deformable Gabor Feature Networks for Biomedical Image Classification

Xuan Gong, Xin Xia, Wentao Zhu, Baochang Zhang, David Doermann, Li'an Zhuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4004-4012

In recent years, deep learning has dominated progress in the field of medical image analysis. We find however, that the ability of current deep learning approaches to represent the complex geometric structures of many medical images is insufficient. One limitation is that deep learning models require a tremendous amount of data, and it is very difficult to obtain a sufficient amount with the necessary detail. A second limitation is that there are underlying features of these medical images that are well established, but the black-box nature of existing convolutional neural networks (CNNs) do not allow us to exploit them. In this paper, we revisit Ga- bor filters and introduce a deformable Gabor convolution (DGConv) to expand deep networks interpretability and enable complex spatial variati ons. The features are learned at deformable sampling locations with adaptive Gabor convolutions to improve representitiveness and robustness to complex objects. The DGConv replaces standard convolutional layers and is easily trained end-to-end, resulting in a deformable Gabor feature network (DGFN) with few additional parameters and minimal additional training cost. We introduce DGFN for addressing deep multi-instance multi-label classification on the INbreast dataset for mammograms and on the ChestX-ray14 dataset for pulmonary x-ray images.

****

Improving Robustness and Uncertainty Modelling in Neural Ordinary Differential Equations

Srinivas Anumasa, P. K. Srijith; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4053-4061

Deep learning models such as Resnets have resulted in state-of-the-art accuracy in many computer vision problems. Neural ordinary differential equations (NODE) provides a continuous depth generalization of Resnets and overcome drawbacks of Resnet such as model selection and parameter complexity. Though NODE is more robust than Resnet, we find that NODE based architectures are still far away from providing robustness and uncertainty handling required for many computer vision problems. We propose novel NODE models which address these drawbacks. In particular, we propose Gaussian processes (GPs) to model the fully connected neural networks in NODE (NODE-GP) to improve robustness and uncertainty handling capabilities of NODE. The proposed model is flexible to accommodate different NODE architectures, and further improves the model selection capabilities in NODEs. We also find that numerical techniques play an important role in modelling NODE robustness, and propose to use different numerical techniques to improve NODE robustness. We demonstrate the superior robustness and uncertainty handling capabilities of proposed models on adversarial attacks and out-of-distribution experiments for the image classification tasks.

****

Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos

Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2363-2372

Taking advantage of human pose data for understanding human activities has attracted much attention these days. However, state-of-the-art pose estimators struggle in obtaining high-quality 2D or 3D pose data due to occlusion, truncation and low-resolution in real-world un-annotated videos. Hence, in this work, we propose 1) a Selective Spatio-Temporal Aggregation mechanism, named SST-A, that refines and smooths the keypoint locations extracted by multiple expert pose estimators, 2) an effective weakly-supervised self-training framework which leverages the aggregated poses as pseudo ground-truth instead of handcrafted annotations for real-world pose estimation. Extensive experiments are conducted for evaluating not only the upstream pose refinement but also the downstream action recognition performance on four datasets, Toyota Smarthome, NTU-RGB+D, Charades, and Kinetics-50. We demonstrate that the skeleton data refined by our Pose-Refinement system (SSTA-PRS) is effective at boosting various existing action recognition models, which achieves competitive or state-of-the-art performance. Refined pose data is available at: https://github.com/walker-a11y/SSTA-PRS

****

Long-Range Attention Network for Multi-View Stereo

Xudong Zhang, Yutao Hu, Haochen Wang, Xianbin Cao, Baochang Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3782-3791

Learning-based multi-view stereo (MVS) has recently gained great popularity, which can efficiently infer depth map and reconstruct fine-grained scene geometry. Previous methods calculate the variance of the corresponding pixel pairs to determine whether they are matched mostly based on the pixel-wise measure, which fails to consider the interdependence among pixels and is ineffective on the matching of texture-less or occluded regions. These false matching problems challenge

MVS and result in its most failure cases. To address the issues, we introduce a Long-range Attention Network (LANet) to selectively aggregate reference features to each position to capture the long-range interdependence across the entire space. As a result, similar features relate to each other regardless of their distance, propagating more guiding information for the effective match. Furthermore, we introduce a new loss to supervise the intermediate probability volume by constraining its distribution reasonably centered at the true depth. Extensive experiments on large-scale DTU dataset demonstrate that the proposed LANet achieves the new state-of-the-art performance, outperforming previous methods by a large margin. Our method is generic and also achieves comparable results on outdoor Tanks and Temples dataset without any fine-tuning, which validates our method's generalization ability.
****
Domain-Adaptive Few-Shot Learning
An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, Ji-Rong Wen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1390-1399
Existing few-shot learning (FSL) methods make the implicit assumption that the few target class samples are from the same domain as the source class samples. However, in practice, this assumption is often invalid -- the target classes could come from a different domain. This poses an additional challenge of domain adaptation (DA) with few training samples. In this paper, the problem of domain-adaptive few-shot learning (DA-FSL) is tackled, which is expected to have wide use in real-world scenarios and requires solving FSL and DA in a unified framework. To this end, we propose a novel domain-adversarial prototypical network (DAPN) model. It is designed to address a specific challenge in DA-FSL: the DA objective means that the source and target data distributions need to be aligned, typically through a shared domain-adaptive feature embedding space; but the FSL objective dictates that the target domain per class distribution must be different from that of any source domain class, meaning aligning the distributions across domains may harm the FSL performance. How to achieve global domain distribution alignment whilst maintaining source/target per-class discriminativeness thus becomes the key. Our solution is to explicitly enhance the source/target per-class separation before domain-adaptive feature embedding learning, to alleviate the negative effect of domain alignment on FSL. Extensive experiments show that our DAPN outperforms the state-of-the-arts. The code is available at https://github.com/dingmyu/DAPN.
****
Lip-Reading With Densely Connected Temporal Convolutional Networks
Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, Maja Pantic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2857-2866
In this work, we present the Densely Connected Temporal Convolutional Network (DC-TCN) for lip-reading of isolated words. Although Temporal Convolutional Networks (TCN) have recently demonstrated great potential in many vision tasks, its receptive fields are not dense enough to model the complex temporal dynamics in lip-reading scenarios. To address this problem, we introduce dense connections into the network to capture more robust temporal features. Moreover, our approach utilises the Squeeze-and-Excitation block, a light-weight attention mechanism, to further enhance the model's classification power. Without bells and whistles, our DC-TCN method has achieved 88.36% accuracy on the Lip Reading in the Wild (LRW) dataset and 43.65% on the LRW-1000 dataset, which has surpassed all the baseline methods and is the new state-of-the-art on both datasets.
****
Auto-Navigator: Decoupled Neural Architecture Search for Visual Navigation
Tianqi Tang, Xin Yu, Xuanyi Dong, Yi Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3743-3752
Existing visual navigation approaches leverage classification neural networks to extract global features from visual data for navigation. However, these networks are not originally designed for navigation tasks. Thus, the neural architectur

es might not be suitable to capture scene contents. Fortunately, neural architecture search (NAS) brings a hope to solve this problem. In this paper, we propose an Auto-Navigator to customize a specialized network for visual navigation. However, as navigation tasks mainly rely on reinforcement learning (RL) rewards in training, such weak supervision is insufficiently indicative for NAS to optimize visual perception network. Thus, we introduce imitation learning (IL) with optimal paths to optimize navigation policies while selecting an optimal architecture. As Auto-Navigator can obtain a direct supervision in every step, such guidance greatly facilitates architecture search. In particular, we initialize our Auto-Navigator with a learnable distribution over the search space of visual perception architecture, and then optimize the distribution with IL supervision. Afterwards, we employ an RL reward function to fine-tune our Auto-Navigator to improve the generalization ability of our model. Extensive experiments demonstrate that our Auto-Navigator outperforms baseline methods on Gibson and Matterport3D without significantly increasing network parameters.

****

Unsupervised Video Representation Learning by Bidirectional Feature Prediction

Nadine Behrmann, Jurgen Gall, Mehdi Noroozi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1670-1679

This paper introduces a novel method for self-supervised video representation learning via feature prediction. In contrast to the previous methods that focus on future feature prediction, we argue that a supervisory signal arising from unobserved past frames is complementary to one that originates from the future frames. The rationale behind our method is to encourage the network to explore the temporal structure of videos by distinguishing between future and past given present observations. We train our model in a contrastive learning framework, where joint encoding of future and past provides us with a comprehensive set of temporal hard negatives via swapping. We empirically show that utilizing both signals enriches the learned representations for the downstream task of action recognition. It outperforms independent prediction of future and past.

****

From Generalized Zero-Shot Learning to Long-Tail With Class Descriptors

Dvir Samuel, Yuval Atzmon, Gal Chechik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 286-295

Real-world data is predominantly unbalanced and long-tailed, but deep models struggle to recognize rare classes in the presence of frequent classes. Often, classes can be accompanied by side information like textual descriptions, but it is not fully clear how to use them for learning with unbalanced long-tail data. Such descriptions have been mostly used in (Generalized) Zero-shot learning (ZSL), suggesting that ZSL with class descriptions may also be useful for long-tail distributions. We describe DRAGON, a late-fusion architecture for long-tail learning with class descriptors. It learns to (1) correct the bias towards head classes on a sample-by-sample basis; and (2) fuse information from class-descriptions to improve the tail-class accuracy. We also introduce new benchmarks CUB-LT, SUN-LT, AWA-LT for long-tail learning with class-descriptions, building on existing learning-with-attributes datasets and a version of Imagenet-LT with class descriptors. DRAGON outperforms state-of-the-art models on the new benchmark. It is also a new SoTA on existing benchmarks for GFSL with class descriptors (GFSL-d) and standard (vision-only) long-tailed learning ImageNet-LT, CIFAR-10, 100, and Places365-LT.

****

CASIA-SURF CeFA: A Benchmark for Multi-Modal Cross-Ethnicity Face Anti-Spoofing

Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, Stan Z. Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1179-1187

The issue of ethnic bias has proven to affect the performance of face recognition in previous works, while it still remains to be vacant in face anti-spoofing. Therefore, in order to study the ethnic bias for face anti-spoofing, we introduce the largest CASIA-SURF Cross-ethnicity Face Anti-spoofing (CeFA) dataset, covering 3 ethnicities, 3 modalities, 1,607 subjects, and 2D plus 3D attack types. F ive protocols are introduced to measure the affect under varied evaluation conditions, such as cross-ethnicity, unknown spoofs or both of them. As our knowledge, CASIA-SURF CeFA is the first dataset including explicit ethnic labels in current released datasets. Then, we propose a novel multi-modal fusion method as a strong baseline to alleviate the ethnic bias, which employs a partially shared fusion strategy to learn complementary information from multiple modalities. Extensive experiments have been conducted on the proposed dataset to verify its significance and generalization capability for other existing datasets, i.e., CASIA-SURF, OULU-NPU and SiW datasets. The dataset is available at https://sites.google.com/qq.com/face-anti spoofing/welcome/challengecvpr2020?authuser=0.

****

Ellipse Detection and Localization With Applications to Knots in Sawn Lumber Images

Shenyi Pan, Shuxian Fan, Samuel W.K. Wong, James V. Zidek, Helge Rhodin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3892-3901

While general object detection has seen tremendous progress, localization of elliptical objects has received little attention in the literature. Our motivating application is the detection of knots in sawn timber images, which is an important problem since the number and types of knots are visual characteristics that adversely affect the quality of sawn timber. We demonstrate how models can be tailored to the elliptical shape and thereby improve on general purpose detectors; more generally, elliptical defects are common in industrial production, such as enclosed air bubbles when casting glass or plastic. In this paper, we adapt the Faster R-CNN with its Region Proposal Network (RPN) to model elliptical objects with a Gaussian function, and extend the existing Gaussian Proposal Network (GPN) architecture by adding the region-of-interest pooling and regression branches, as well as using the Wasserstein distance as the loss function to predict the precise locations of elliptical objects. Our proposed method has promising results on the lumber knot dataset: knots are detected with an average intersection over union of 73.05%, compared to 63.63% for general purpose detectors. Specific to the lumber application, we also propose an algorithm to correct any misalignment in the raw timber images during scanning, and contribute the first open-source lumber knot dataset by labeling the elliptical knots in the preprocessed images.

****

Foreground-Aware Semantic Representations for Image Harmonization

Konstantin Sofiiuk, Polina Popenova, Anton Konushin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1620-1629

Image harmonization is an important step in photo editing to achieve visual consistency in composite images by adjusting the appearances of a foreground to make it compatible with a background. Previous approaches to harmonize composites are based on training of encoder-decoder networks from scratch, which makes it challenging for a neural network to learn a high-level representation of objects. We propose a novel architecture to utilize the space of high-level features learned by a pre-trained classification network. We create our models as a combination of existing encoder-decoder architectures and a pre-trained foreground-aware deep high-resolution network. We extensively evaluate the proposed method on existing image harmonization benchmark and set up a new state-of-the-art in terms of MSE and PSNR metrics.

****

Action Duration Prediction for Segment-Level Alignment of Weakly-Labeled Videos

Reza Ghoddoosian, Saif Sayed, Vassilis Athitsos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2053-2062

This paper focuses on weakly-supervised action alignment, where only the ordered sequence of video-level actions is available for training. We propose a novel Duration Network, which captures a short temporal window of the video and learns to predict the remaining duration of a given action at any point in time with a level of granularity based on the type of that action. Further, we introduce a S egment-Level Beam Search to obtain the best alignment, that maximizes our posterior probability. Segment-Level Beam Search efficiently aligns actions by considering only a selected set of frames that have more confident predictions. The experimental results show that our alignments for long videos are more robust than existing models. Moreover, the proposed method achieves state of the art results in certain cases on the popular Breakfast and Hollywood Extended datasets.

****

DACS: Domain Adaptation via Cross-Domain Mixed Sampling

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, Lennart Svensson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1379-1389

Semantic segmentation models based on convolutional neural networks have recently displayed remarkable performance for a multitude of applications. However, these models typically do not generalize well when applied on new domains, especially when going from synthetic to real data. In this paper we address the problem of unsupervised domain adaptation (UDA), which attempts to train on labelled data from one domain (source domain), and simultaneously learn from unlabelled data in the domain of interest (target domain). Existing methods have seen success by training on pseudo-labels for these unlabelled images. Multiple techniques have been proposed to mitigate low-quality pseudo-labels arising from the domain shift, with varying degrees of success. We propose DACS: Domain Adaptation via Cross-domain mixed Sampling, which mixes images from the two domains along with the corresponding labels and pseudo-labels. These mixed samples are then trained on, in addition to the labelled data itself. We demonstrate the effectiveness of our solution by achieving state-of-the-art results for GTA5 to Cityscapes, a common synthetic-to-real semantic segmentation benchmark for UDA.

****

SliceNets -- A Scalable Approach for Object Detection in 3D CT Scans

Anqi Yang, Feng Pan, Vishwanath Saragadam, Duy Dao, Zhuo Hui, Jen-Hao Rick Chang, Aswin C. Sankaranarayanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 335-344

One of the most promising approaches for automated detection of guns and other prohibited items in aviation baggage screening is the use of 3D computed tomography (CT) scans. However, automated detection, especially with deep neural networks, faces two key challenges: the high dimensionality of individual 3D scans, and the lack of labeled training data. We address these challenges using a novel image-based detection and segmentation technique that we call the slice-and-fuse framework. Our approach relies on slicing the input 3D volumes along the three cardinal directions, generating 2D predictions on each slice using 2D CNNs, and subsequently fusing the 2D predictions to obtain a 3D prediction. We develop two distinct detectors based on this slice-and-fuse strategy: the Retinal-SliceNet that uses a unified, single network with end-to-end training, and the U-SliceNet that uses a two-stage paradigm, first generating proposals using a voxel labeling network and, subsequently, refining the proposals by a 3D classification network. The networks are trained using a data augmentation approach that creates a very large training dataset by inserting weapons into 3D CT scans of threat-free bags. We demonstrate that the two SliceNets outperform state-of-the-art 3D object detection methods on a large-scale 3D baggage CT dataset for baggage classification, 3D object detection, and 3D semantic segmentation.

****

Real-Time Localized Photorealistic Video Style Transfer

Xide Xia, Tianfan Xue, Wei-Sheng Lai, Zheng Sun, Abby Chang, Brian Kulis, Jiawen Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1089-1098

We present a novel algorithm for transferring artistic styles of semantically meaningful local regions of an image onto local regions of a target video while preserving its photorealism. Local regions may be selected either fully automatically from an image, through using video segmentation algorithms, or from casual user guidance such as scribbles. Our method, based on a deep neural network architecture inspired by recent work in photorealistic style transfer, is real-time a

nd works on arbitrary inputs without runtime optimization once trained on a diverse dataset of artistic styles. By augmenting our video dataset with noisy semantic labels and jointly optimizing over style, content, mask, and temporal losses, our method can cope with a variety of imperfections in the input and produce temporally coherent videos without visual artifacts. We demonstrate our method on a variety of style images and target videos, including the ability to transfer different styles onto multiple objects simultaneously, and smoothly transition between styles in time.
****

CoMoDA: Continuous Monocular Depth Adaptation Using Past Experiences
Yevhen Kuznietsov, Marc Proesmans, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2907-2917
While ground truth depth data remains hard to obtain, self-supervised monocular depth estimation methods enjoy growing attention. Much research in this area aims at improving loss functions or network architectures. Most works, however, do not leverage self-supervision to its full potential. They stick to the standard closed world train-test pipeline, assuming the network parameters to be fixed after the training is finished. Such an assumption does not allow to adapt to new scenes, whereas with self-supervision this becomes possible without extra annotations. In this paper, we propose a novel self-supervised Continuous Monocular Depth Adaptation method (CoMoDA), which adapts the pretrained model on a test video on the fly. As opposed to existing test-time refinement methods that use isolated frame triplets, we opt for continuous adaptation, making use of the previous experience from the same scene. We additionally augment the proposed procedure with the experience from the distant past, preventing the model from overfitting and thus forgetting already learnt information. We demonstrate that our method can be used for both intra- and cross-dataset adaptation. By adapting the model from train to test set of the Eigen split of KITTI, we achieve state-of-the-art depth estimation performance and surpass all existing methods using standard architectures. We also show that our method runs 15 times faster than existing test-time refinement methods. The code is available at https://github.com/Yevkuzn/CoMoDA.
****

HyperCon: Image-to-Video Model Transfer for Video-to-Video Translation Tasks
Ryan Szeto, Mostafa El-Khamy, Jungwon Lee, Jason J. Corso; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3080-3089
Video-to-video translation is more difficult than image-to-image translation due to the temporal consistency problem that, if unaddressed, leads to distracting flickering effects. Although video models designed from scratch produce temporally consistent results, training them to match the vast visual knowledge captured by image models requires an intractable number of videos. To combine the benefits of image and video models, we propose an image-to-video model transfer method called Hyperconsistency (HyperCon) that transforms any well-trained image model into a temporally consistent video model without fine-tuning. HyperCon works by translating a temporally interpolated video frame-wise and then aggregating over temporally localized windows on the interpolated video. It handles both masked and unmasked inputs, enabling support for even more video-to-video translation tasks than prior image-to-video model transfer techniques. We demonstrate HyperCon on video style transfer and inpainting, where it performs favorably compared to prior state-of-the-art methods without training on a single stylized or incomplete video.
****

Improving Video Captioning With Temporal Composition of a Visual-Syntactic Embedding
Jesus Perez-Martin, Benjamin Bustos, Jorge Perez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3039-3049
Video captioning is the task of predicting a semantic and syntactically correct sequence of words given some context video. The most successful methods for video captioning have a strong dependency on the effectiveness of semantic represent

ations learned from visual models, but often produce syntactically incorrect sentences which harms their performance on standard datasets. In this paper, we address this limitation by considering syntactic representation learning as an essential component of video captioning. We construct a visual-syntactic embedding by mapping into a common vector space a visual representation, that depends only on the video, with a syntactic representation that depends only on Part-of-Speech (POS) tagging structures of the video description. We integrate this joint representation into an encoder-decoder architecture that we call Visual-Semantic-Syntactic Aligned Network (SemSynAN), which guides the decoder (text generation stage) by aligning temporal compositions of visual, semantic, and syntactic representations. We tested our proposed architecture obtaining state-of-the-art results on two widely used video captioning datasets: the Microsoft Video Description (MSVD) dataset and the Microsoft Research Video-to-Text (MSR-VTT) dataset.

****

Adaptive Multiplane Image Generation From a Single Internet Picture

Diogo C. Luvizon, Gustavo Sutter P. Carvalho, Andreza A. dos Santos, Jhonatas S. Conceicao, Jose L. Flores-Campana, Luis G. L. Decker, Marcos R. Souza, Helio Pedrini, Antonio Joia, Otavio A. B. Penatti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2556-2565

In the last few years, several works have tackled the problem of novel view synthesis from a pair of stereo images or even from a single picture. However, previous methods are computationally expensive, specially for high-resolution images. In this paper, we address the problem of generating an efficient multiplane image (MPI) from a single high-resolution picture. We present the adaptive-MPI representation, which allows rendering novel views with low computational requirements. To this end, we propose an adaptive slicing algorithm that produces an MPI with a variable number of image planes. We also present a new lightweight CNN for depth estimation, which is learned by knowledge distillation from a larger network. Occluded regions in the adaptive-MPI are inpainted also by a lightweight CNN. We show that our method is capable of producing high-quality predictions with one order of magnitude less parameters, when compared to previous approaches. In addition, we show the robustness of our method for novel view synthesis on challenging pictures from the Internet.

****

MSNet: A Multilevel Instance Segmentation Network for Natural Disaster Damage Assessment in Aerial Videos

Xiaoyu Zhu, Junwei Liang, Alexander Hauptmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2023-2032

In this paper, we study the problem of efficiently assessing building damage after natural disasters like hurricanes, floods or fires, through aerial video analysis. We make two main contributions. The first contribution is a new dataset, consisting of user-generated aerial videos from social media with annotations of instance-level building damage masks. This provides the first benchmark for quantitative evaluation of models to assess building damage using aerial videos. The second contribution is a new model, namely MSNet, which contains novel region proposal network designs and an unsupervised score refinement network for confidence score calibration in both bounding box and mask branches. We show that our model achieves state-of-the-art results compared to previous methods in our dataset.

****

Recovering Trajectories of Unmarked Joints in 3D Human Actions Using Latent Space Optimization

Suhas Lohit, Rushil Anirudh, Pavan Turaga; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2342-2351

Motion capture (mocap) and time-of-flight based sensing of human actions are becoming increasingly popular modalities to perform robust activity analysis. Applications range from action recognition to quantifying movement quality for health applications. While marker-less motion capture has made great progress, in critical applications such as healthcare, marker-based systems, especially active markers, are still considered gold-standard. However, there are several practical challenges in both modalities such as visibility, tracking errors, and simply the need to keep marker setup convenient wherein movements are recorded with a reduced marker-set. This implies that certain joint locations will not even be marked-up, making downstream analysis of full body movement challenging. To address this gap, we first pose the problem of reconstructing the unmarked joint data as an ill-posed linear inverse problem. We recover missing joints for a given action by projecting it onto the manifold of human actions, this is achieved by optimizing the latent space representation of a deep autoencoder. Experiments on both mocap and Kinect datasets clearly demonstrate that the proposed method performs very well in recovering semantics of the actions and dynamics of missing joints. We will release all the code and models publicly.

****

Do Not Forget to Attend to Uncertainty While Mitigating Catastrophic Forgetting

Vinod K. Kurmi, Badri N. Patro, Venkatesh K. Subramanian, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 736-745

One of the major limitations of deep learning models is that they face catastrophic forgetting in an incremental learning scenario. There have been several approaches proposed to tackle the problem of incremental learning. Most of these methods are based on knowledge distillation and do not adequately utilize the information provided by older task models, such as uncertainty estimation in predictions. The predictive uncertainty provides the distributional information can be applied to mitigate catastrophic forgetting in a deep learning framework. In the proposed work, we consider a Bayesian formulation to obtain the data and model uncertainties. We also incorporate self-attention framework to address the incremental learning problem. We define distillation losses in terms of aleatoric uncertainty and self-attention. In the proposed work, we investigate different ablation analyses on these losses. Furthermore, we are able to obtain better results in terms of accuracy on standard benchmarks.

****

ResNet or DenseNet? Introducing Dense Shortcuts to ResNet

Chaoning Zhang, Philipp Benz, Dawit Mureja Argaw, Seokju Lee, Junsik Kim, Francois Rameau, Jean-Charles Bazin, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3550-3559

ResNet or DenseNet? Nowadays, most deep learning based approaches are implemented with seminal backbone networks, among them the two arguably most famous ones are ResNet and DenseNet. Despite their competitive performance and overwhelming popularity, inherent drawbacks exist for both of them. For ResNet, the identity shortcut that stabilizes training might limit its representation capacity, and DenseNet mitigates it with multi-layer feature concatenation. However, the dense concatenation causes a new problem of requiring high GPU memory and more training time. Partially due to this, it is not a trivial choice between ResNet and DenseNet. This paper provides a unified perspective of dense summation to analyze them, which facilitates a better understanding of their core difference. We further propose dense weighted normalized shortcuts as a solution to the dilemma between them. Our proposed dense shortcut inherits the design philosophy of simple design in ResNet and DenseNet. On several benchmark datasets, the experimental results show that the proposed DSNet achieves significantly better results than ResNet, and achieves comparable performance as DenseNet but requiring fewer computation resources.

****

Multi-Task Knowledge Distillation for Eye Disease Prediction

Sahil Chelaramani, Manish Gupta, Vipul Agarwal, Prashant Gupta, Ranya Habash; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3983-3993

While accurate disease prediction from retinal fundus images is critical, collecting large amounts of high quality labeled training data to build such supervised models is difficult. Deep learning classifiers have led to high accuracy results across a wide variety of medical imaging problems, but they need large amounts of labeled data. Given a fundus image, we aim to evaluate various solutions fo r learning deep neural classifiers using small labeled data for three tasks related to eye disease prediction: (T1) predicting one of the five broad categories - diabetic retinopathy, age-related macular degeneration, glaucoma, melanoma and normal, (T2) predicting one of the 320 fine-grained disease sub-categories, (T3) generating a textual diagnosis. The problem is challenging because of small data size, need for predictions across multiple tasks, handling image variations, and large number of hyper-parameter choices. Modeling the problem under a multi-task learning (MTL) setup, we investigate the contributions of each of the proposed tasks while dealing with a small amount of labeled data. Further, we suggest a novel MTL-based teacher ensemble method for knowledge distillation. On a data set of 7212 labeled and 35854 unlabeled images across 3502 patients, our technique obtains 83% accuracy, 75% top-5 accuracy and 48 BLEU for tasks T1, T2 and T3 respectively. Even with 15% training data, our method outperforms baselines by 8.1, 3.2 and 11.2 points for the three tasks respectively.

****

SALAD: Self-Assessment Learning for Action Detection

Guillaume Vaudaux-Ruth, Adrien Chan-Hon-Tong, Catherine Achard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1269-1278

Literature on self-assessment in machine learning mainly focuses on the production of well-calibrated algorithms through consensus frameworks i.e. calibration is seen as a problem. Yet, we observe that learning to be properly confident could behave like a powerful regularization and thus, could be an opportunity to improve performance. Precisely, we show that used within a framework of action detection, the learning of a self-assessment score is able to improve the whole action localization process. Experimental results show that our approach outperforms the state-of-the-art on two action detection benchmarks. On THUMOS14 dataset, the mAP at tIoU@0.5 is improved from 42.8% to 44.6%, and from 50.4% to 51.7% on ActivityNet1.3 dataset. For lower tIoU values, we achieve even more significant improvements on both datasets.

****

Dense-Resolution Network for Point Cloud Classification and Segmentation

Shi Qiu, Saeed Anwar, Nick Barnes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3813-3822

Point cloud analysis is attracting attention from Artificial Intelligence research since it can be widely used in applications such as robotics, Augmented Reality, self-driving. However, it is always challenging due to irregularities, unorderedness, and sparsity. In this article, we propose a novel network named Dense-Resolution Network (DRNet) for point cloud analysis. Our DRNet is designed to learn local point features from the point cloud in different resolutions. In order to learn local point groups more effectively, we present a novel grouping method for local neighborhood searching and an error-minimizing module for capturing local features. In addition to validating the network on widely used point cloud segmentation and classification benchmarks, we also test and visualize the performance of the components. Comparing with other state-of-the-art methods, our network shows superiority on ModelNet40, ShapeNet synthetic and ScanObjectNN real point cloud datasets.

****

Distillation Multiple Choice Learning for Multimodal Action Recognition

Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, Stan Sclaroff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2755-2764

In this work, we address the problem of learning an ensemble of specialist networks using multimodal data, while considering the realistic and challenging scenario of possible missing modalities at test time. Our goal is to leverage the complementary information of multiple modalities to the benefit of the ensemble and each individual network. We introduce a novel Distillation Multiple Choice Learning framework for multimodal data, where different modality networks learn in a cooperative setting from scratch, strengthening one another. The modality networks learned using our method achieve significantly higher accuracy than if train ed separately, due to the guidance of other modalities. We evaluate this approach on three video action recognition benchmark datasets. We obtain state-of-the-art results in comparison to other approaches that work with missing modalities at test time.

****

Unsupervised Meta-Domain Adaptation for Fashion Retrieval

Vivek Sharma, Naila Murray, Diane Larlus, Saquib Sarfraz, Rainer Stiefelhagen, Gabriela Csurka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1348-1357

Unsupervised Meta-Domain Adaptation for Fashion RetrievalCross-domain fashion item retrieval naturally arises when unconstrained consumer images are used to query for fashion items in a collection of high-end photographs provided by retailers. To perform this task, approaches typically leverage both consumer and shop domains from a given dataset to learn a domain invariant representation, allowing these images of different nature to be directly compared. When consumer images are not available beforehand, such training is impossible. In this paper, we focus on this challenging and yet practical scenario, and we propose instead to leverage representations learned for cross-domain retrieval from another source dataset and to adapt them to the target dataset for this particular setting. More precisely, we bypass the lack of consumer images and directly target the more challenging meta-domain gap which occurs between consumer images and shop images, independently of their dataset. Assuming that datasets share some similar fashion items, we cluster their shop images and leverage the clusters to automatically generate pseudo-labels. Those are used to associate consumer and shop images across datasets, which in turn allows to learn meta-domain-invariant representations suitable for cross-domain retrieval in the target dataset.

****

SChISM: Semantic Clustering via Image Sequence Merging for Images of Human-Decomposition

Sara Mousavi, Dylan Lee, Tatianna Griffin, Kelley Cross, Dawnie Steadman, Audris Mockus; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2190-2199

In many domains, large image collections are key ways in which information about relevant phenomena is retained and analyzed, yet it remains challenging to use such data in research and practice. Our aim is to investigate this problem in the context of a forensic unlabeled dataset of over 1M human decomposition photos. To make this collection usable by experts, various body parts first need to be identified and traced through their evolution despite their distinct appearances at different stages of decay from "fresh" to "skeletonized". We developed an unsupervised technique for clustering images that builds sequences of similar images representing the evolution of each body part through stages of decomposition. Evaluation of our method on 34,476 human decomposition images shows that our method significantly outperforms the state of the art clustering method in this application.

****

Active Latent Space Shape Model: A Bayesian Treatment of Shape Model Adaptation With an Application to Psoriatic Arthritis Radiographs

Adwaye Rambojun, William Tillett, Tony Shardlow, Neill D. F. Campbell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2043-2052

Shape models have been used extensively to regularise segmentation of objects of interest in images, e.g. bones in medical x-ray radiographs, given supervised training examples. However, approaches usually adopt simple linear models that do not capture uncertainty and require extensive annotation effort to label a large number of set template landmarks for training. Conversely, supervised deep learning methods have been used on appearance directly (no explicit shape modelling) but these fail to capture detailed features that are clinically important. We present a supervised approach that combines both a non-linear generative shape model and a discriminative appearance-based convolutional neural network whilst quantifying uncertainty and relaxes the need for detailed, template based alignme nt for the training data. Our Bayesian framework couples the uncertainty from both the generator and the discriminator; our main contribution is the marginalisation of an intractable integral through the use of radial basis function approximations. We illustrate this model on the problem of segmenting bones from Psoriatic Arthritis hand radiographs and demonstrate that we can accurately measure the clinically important joint space gap between neighbouring bones.

****

Conditional Link Prediction of Category-Implicit Keypoint Detection

Ellen Yi-Ge, Rui Fan, Zechun Liu, Zhiqiang Shen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3440-3449

Keypoints of objects reflect their concise abstractions, while the corresponding connection links (CL) build the skeleton by detecting the intrinsic relations between keypoints. Existing approaches are typically computationally-intensive, inapplicable for instances belonging to multiple classes, and/or infeasible to simultaneously encode connection information. To address the aforementioned issues, we propose an end-to-end category-implicit Keypoint and Link Prediction Network (KLPNet), which is the first approach for simultaneous semantic keypoint detection (for multi-class instances) and CL rejuvenation. In our KLPNet, a novel Conditional Link Prediction Graph is proposed for link prediction among keypoints that are contingent on a predefined category. Furthermore, a Cross-stage Keypoint Localization Module (CKLM) is introduced to explore feature aggregation for coarse-to-fine keypoint localization. Comprehensive experiments conducted on three publicly available benchmarks demonstrate that our KLPNet consistently outperforms all other state-of-the-art approaches. Furthermore, the experimental results of CL prediction also show the effectiveness of our KLPNet with respect to occlusion problems.

****

Noise as a Resource for Learning in Knowledge Distillation

Elahe Arani, Fahad Sarfraz, Bahram Zonooz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3129-3138

While noise is commonly considered a nuisance in computing systems, a number of studies in neuroscience have shown several benefits of noise in the nervous system from enabling the brain to carry out computations such as probabilistic inference as well as carrying additional information about the stimuli. Similarly, noise has been shown to improve the performance of deep neural networks. In this study, we further investigate the effect of adding noise in the knowledge distillation framework because of its resemblance to collaborative subnetworks in the brain regions. We empirically show that injecting constructive noise at different levels in the collaborative learning framework enables us to train the model effectively and distill desirable characteristics in the student model. In doing so, we propose three different methods that target the common challenges in deep neural networks: minimizing the performance gap between a compact model and large model (Fickle Teacher), training high performance compact adversarially robust models (Soft Randomization), and training models efficiently under label noise (Messy Collaboration). Our findings motivate further study in the role of noise as a resource for learning in a collaborative learning framework.

****

Real-Time Uncertainty Estimation in Computer Vision via Uncertainty-Aware Distribution Distillation

Yichen Shen, Zhilu Zhang, Mert R. Sabuncu, Lin Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 707-716

Calibrated estimates of uncertainty are critical for many real-world computer vision applications of deep learning. While there are several widely-used uncertainty estimation methods, dropout inference stands out for its simplicity and efficacy. This technique, however, requires multiple forward passes through the network during inference and therefore can be too resource-intensive to be deployed in real-time applications. To tackle this issue, we propose a unified distillation paradigm for learning the conditional predictive distribution of a pre-trained dropout model for fast uncertainty estimation of both aleatoric and epistemic uncertainty at the same time. We empirically test the effectiveness of the propo sed method on both semantic segmentation and depth estimation tasks and observe that the student model can well approximate the probability distribution generated by the teacher model, i.e the pre-trained dropout model. In addition to a significant boost in speed, we demonstrate the quality of uncertainty estimates and the overall predictive performance can also be improved with the proposed method.

****

Constrained Weight Optimization for Learning Without Activation Normalization

Daiki Ikami, Go Irie, Takashi Shibata; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2606-2614

Weight Normalization (WN) is an essential building block in deep learning. However, even state-of-the-art WN methods need to be combined with activation normalization methods, such as Batch Normalization (BN), to provide the same classification accuracy as BN. In this paper, we aim to circumvent this issue with a weight normalization approach that can be used on its own to provide a classification accuracy competitive to BN. Our approach mimics three fundamental properties of BN, namely, keeping the norm of the weights constant, setting the mean of the weights to zero, and simulating stochastic perturbations due to batch sampling bias. Unlike most of the existing WN methods that rely on "reparametrization", our method directly optimizes the weights with proper constraints and thus can circumvent its serious drawback, gradient explosion. Moreover, we propose an efficient and easy-to-implement algorithm to solve our constrained optimization problem without sacrificing its benefits. The results of classification experiments on three popular benchmark datasets demonstrate that our method is highly competitive with or even better than the state-of-the-art normalization methods.

****

Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection

Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, Shijian Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2524-2534

Automated defect inspection is critical for effective and efficient maintenance, repair, and operations in advanced manufacturing. On the other hand, automated defect inspection is often constrained by the lack of defect samples, especially when we adopt deep neural networks for this task. This paper presents Defect-GAN, an automated defect synthesis network that generates realistic and diverse defect samples for training accurate and robust defect inspection networks. Defect-GAN learns through defacement and restoration processes, where the defacement generates defects on normal surface images while the restoration removes defects to generate normal images. It employs a novel compositional layer-based architecture for generating realistic defects within various image backgrounds with different textures and appearances. It can also mimic the stochastic variations of defects and offer flexible control over the locations and categories of the generated defects within the image background. Extensive experiments show that Defect-GAN is capable of synthesizing various defects with superior diversity and fidelity. In addition, the synthesized defect samples demonstrate their effectiveness in training better defect inspection networks.

****

Set Augmented Triplet Loss for Video Person Re-Identification

Pengfei Fang, Pan Ji, Lars Petersson, Mehrtash Harandi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 464-473

Modern video person re-identification (re-ID) machines are often trained using a metric learning approach, supervised by a triplet loss. The triplet loss used in video re-ID is usually based on so-called clip features, each aggregated from a few frame features. In this paper, we propose to model the video clip as a set and instead study the distance between sets in the corresponding triplet loss. In contrast to the distance between clip representations, the distance between clip sets considers the pair-wise similarity of each element (i.e., frame representation) between two sets. This allows the network to directly optimize the feature representation at a frame level. Apart from the commonly-used set distance metrics (e.g., ordinary distance and Hausdorff distance), we further propose a hy

brid distance metric, tailored for the set-aware triplet loss. Also, we propose a hard positive set construction strategy using the learned class prototypes in a batch. Our proposed method achieves state-of-the-art results across several standard benchmarks, demonstrating the advantages of the proposed method.
****

A Deep Temporal Fusion Framework for Scene Flow Using a Learnable Motion Model and Occlusions
Rene Schuster, Christian Unger, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 247-255
Motion estimation is one of the core challenges in computer vision. With traditional dual-frame approaches, occlusions and out-of-view motions are a limiting factor, especially in the context of environmental perception for vehicles due to the large (ego-) motion of objects. Our work proposes a novel data-driven approach for temporal fusion of scene flow estimates in a multi-frame setup to overcome the issue of occlusion. Contrary to most previous methods, we do not rely on a constant motion model, but instead learn a generic temporal relation of motion from data. In a second step, a neural network combines bi-directional scene flow estimates from a common reference frame, yielding a refined estimate and a natural byproduct of occlusion masks. This way, our approach provides a fast multi-frame extension for a variety of scene flow estimators, which outperforms the underlying dual-frame approaches.
****

MART: Motion-Aware Recurrent Neural Network for Robust Visual Tracking
Heng Fan, Haibin Ling; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 566-575
We introduce MART, Motion-Aware Recurrent neural network (MA-RNN) for Tracking, by modeling robust long-term spatial-temporal representation. In particular, we propose a simple, yet effective context-aware displacement attention (CADA) module to capture target motion in videos. By seamlessly integrating CADA into RNN, the proposed MA-RNN can spatially align and aggregate temporal information guided by motion from frame to frame, leading to more effective representation that benefits a tracker from motion when handling occlusion, deformation, viewpoint change etc. Moreover, to deal with scale change, we present a monotonic bounding box regression (mBBR) approach that iteratively predicts regression offsets for target object under the guidance of intersection-over-union (IoU) score, guaranteeing non-decreasing accuracy. In extensive experiments on five benchmarks, including GOT-10k, LaSOT, TC-128, OTB-15 and VOT-19, our tracker MART consistently achieves state-of-the-art results and runs in real-time.
****

Splatty- a Unified Image Demosaicing and Rectification Method
Pranav Verma, Dominique E. Meyer, Hanyang Xu, Falko Kuester; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 786-795
Image demosaicing and rectification are key tasks that are frequently used in many computer vision systems. To date, however, their implementations have been plagued with large memory requirements and inconvenient dataflow, making it difficult to scale them to real-time, high resolution settings. This has motivated the development of joint demosaicing and rectification algorithms that resolve the backward mapping dataflow for improved hardware implementation. Towards this purpose, we propose Splatty: an algorithmic solution to pipelined image stream demosaicing and rectification for memory bound applications requiring computational efficiency. We begin by introducing a polynomial Look-up-Table (LUT) compression scheme that can encode any arbitrarily complex lens model for rectification while keeping the remapping errors below 1E-10 pixels, and reducing the memory footprint to O(min(m,n)) from O(mn) for an mxn sized image. The core contribution leverages this LUT for a unified, forward-only splatting algorithm for simultaneous demosaicing and rectification. We demonstrate that merging these two steps into a single, forward-only splatting pass with interpolation, provides distinctive dataflow and performance efficiency benefits while maintaining quality standards when compared to state-of-the-art demosaicing and rectification algorithms.

****

End-to-End Lane Shape Prediction With Transformers

Ruijin Liu, Zejian Yuan, Tie Liu, Zhiliang Xiong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3694-3702

Lane detection, the process of identifying lane markings as approximated curves, is widely used for lane departure warning and adaptive cruise control in autonomous vehicles. The popular pipeline that solves it in two steps---feature extraction plus post-processing, while useful, is too inefficient and flawed in learning the global context and lanes' long and thin structures. To tackle these issues, we propose an end-to-end method that directly outputs parameters of a lane shape model, using a network built with a transformer to learn richer structures and context. The lane shape model is formulated based on road structures and camera pose, providing physical interpretation for parameters of network output. The transformer models non-local interactions with a self-attention mechanism to capture slender structures and global context. The proposed method is validated on the TuSimple benchmark and shows state-of-the-art accuracy with the most lightweight model size and fastest speed. Additionally, our method shows excellent adaptability to a challenging self-collected lane detection dataset, showing its powerful deployment potential in real applications. Codes are available at https://github.com/liuruijin17/E2ELSPTRs.

****

Coarse- and Fine-Grained Attention Network With Background-Aware Loss for Crowd Density Map Estimation

Liangzi Rong, Chunping Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3675-3684

In this paper, we present a novel method Coarse- and Fine-grained Attention Network (CFANet) for generating high-quality crowd density maps and people count estimation by incorporating attention maps to better focus on the crowd area. We devise a from-coarse-to-fine progressive attention mechanism by integrating Crowd Region Recognizer (CRR) and Density Level Estimator (DLE) branch, which can suppress the influence of irrelevant background and assign attention weights according to the crowd density levels, because generating accurate fine-grained attention maps directly is normally difficult. We also employ a multi-level supervision mechanism to assist the backpropagation of gradient and reduce overfitting. Besides, we propose a Background-aware Structural Loss (BSL) to reduce the false recognition ratio while improving the structural similarity to groundtruth. Extensive experiments on commonly used datasets show that our method can not only outperform previous state-of-the-art methods in terms of count accuracy but also improve the image quality of density maps as well as reduce the false recognition ratio.

****

Ensembling Low Precision Models for Binary Biomedical Image Segmentation

Tianyu Ma, Hang Zhang, Hanley Ong, Amar Vora, Thanh D. Nguyen, Ajay Gupta, Yi Wang, Mert R. Sabuncu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 325-334

Segmentation of anatomical regions of interest such as vessels or small lesions in medical images is still a difficult problem that is often tackled with manual input by an expert. One of the major challenges for this task is that the appearance of foreground (positive) regions can be similar to background (negative) regions. As a result, many automatic segmentation algorithms tend to exhibit asymmetric errors, typically producing more false positives than false negatives. In this paper, we aim to leverage this asymmetry and train a diverse ensemble of models with very high recall, while sacrificing their precision. Our core idea is straightforward: A diverse ensemble of low precision and high recall models are likely to make different false positive errors (classifying background as foreground in different parts of the image), but the true positives will tend to be consistent. Thus, in aggregate the false positive errors will cancel out, yielding high performance for the ensemble. Our strategy is general and can be applied with any segmentation model. In three different applications (carotid artery segmentation in a neck CT angiography, myocardium segmentation in a cardiovascular

MRI and multiple sclerosis lesion segmentation in a brain MRI), we show how the proposed approach can significantly boost the performance of a baseline segmentation method.

****

Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes

Loc Trinh, Michael Tsang, Sirisha Rambhatla, Yan Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1973-1983

In this paper we propose a novel human-centered approach for detecting forgery in face images, using dynamic prototypes as a form of visual explanations. Currently, most state-of-the-art deepfake detections are based on black-box models that process videos frame-by-frame for inference, and few closely examine their temporal inconsistencies. However, the existence of such temporal artifacts within deepfake videos is key in detecting and explaining deepfakes to a supervising human. To this end, we propose Dynamic Prototype Network (DPNet) -- an interpretable and effective solution that utilizes dynamic representations (i.e., prototypes) to explain deepfake temporal artifacts. Extensive experimental results show that DPNet achieves competitive predictive performance, even on unseen testing datasets such as Google's DeepFakeDetection, DeeperForensics, and Celeb-DF, while providing easy referential explanations of deepfake dynamics. On top of DPNet's prototypical framework, we further formulate temporal logic specifications based on these dynamics to check our model's compliance to desired temporal behaviors, hence providing trustworthiness for such critical detection systems.

****

TranstextNet: Transducing Text for Recognizing Unseen Visual Relationships

Gal S. Kenigsfield, Ran El-Yaniv; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1955-1964

An important challenge in visual scene understanding is the recognition of interactions between objects in an image. This task - often called visual relationship detection (VRD) - must be solved to enable higher understanding of the semantic content in images. VRD can become particularly hard where there is severe statistical sparsity of some potentially involved objects, and the number of many relationships in standard training sets is limited. In this paper we show how to transduce auxiliary text so as to enable recognition of relationships absent in the visual training data. This transduction is performed by learning a shared relationship representation for both the textual and visual information. The proposed approach is model-agnostic and can be used as a plug-in module in existing VRD and scene graph generation (SGG) recognition systems to improve their performance and extend their capabilities. We consider the application of our technique using three widely accepted SGG models [20, 24, 16], and different auxiliary text sources: image captions, text generated by a deep text generation model (GPT-2), and ebooks from the Gutenberg Project. We conduct an extensive empirical study of both the VRD and SGG tasks over large-scale benchmark datasets. Our method is the first to enable recognition of visual relationships missing in the visual training data and appearing only in the auxiliary text. We conclusively show that text ingestion enables recognition of unseen visual relationships, and moreover, advances the state-of-the-art in all SGG tasks.

****

3D Dense Geometry-Guided Facial Expression Synthesis by Adversarial Learning

Rumeysa Bodur, Binod Bhattarai, Tae-Kyun Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2392-2401

Manipulating facial expressions is a challenging task due to fine-grained shape changes produced by facial muscles and the lack of input-output pairs for supervised learning. Unlike previous methods using Generative Adversarial Networks (GAN), which rely on cycle-consistency loss or sparse geometry (landmarks) loss for expression synthesis, we propose a novel GAN framework to exploit 3D dense (depth and surface normals) information for expression manipulation. However, a large-scale dataset containing RGB images with expression annotations and their corresponding depth maps is not available. To this end, we propose to use an off-the-shelf state-of-the-art 3D reconstruction model to estimate the depth and create a large-scale RGB-Depth dataset after a manual data clean-up process. We utilise this dataset to minimise the novel depth consistency loss via adversarial learning (note we do not have ground truth depth maps for generated face images) and the depth categorical loss of synthetic data on the discriminator. In addition, to improve the generalisation and lower the bias of the depth parameters, we propose to use a novel confidence regulariser on the discriminator side of the framework. We extensively performed both quantitative and qualitative evaluations on two publicly available challenging facial expression benchmarks: AffectNet and RaFD. Our experiments demonstrate that the proposed method outperforms the competitive baseline and existing arts by a large margin.

****

MPRNet: Multi-Path Residual Network for Lightweight Image Super Resolution

Armin Mehri, Parichehr B. Ardakani, Angel D. Sappa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2704-2713

Lightweight super resolution networks have extremely importance for real-world applications. In recent years several SR deep learning approaches with outstanding achievement have been introduced by sacrificing memory and computational cost. To overcome this problem, a novel lightweight super resolution network is proposed, which improves the SOTA performance in lightweight SR and performs roughly similar to computationally expensive networks. Multi-Path Residual Network designs with a set of Residual concatenation Blocks stacked with Adaptive Residual Blocks: (i) to adaptively extract informative features and learn more expressive spatial context information; (ii) to better leverage multi-level representations before up-sampling stage; and (iii) to allow an efficient information and gradient flow within the network. The proposed architecture also contains a new attention mechanism, Two-Fold Attention Module, to maximize the representation ability of the model. Extensive experiments show the superiority of our model against other SOTA SR approaches.

****

A Multi-Class Hinge Loss for Conditional GANs

Ilya Kavalerov, Wojciech Czaja, Rama Chellappa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1290-1299

We propose a new algorithm to incorporate class conditional information into the critic of GANs via a multi-class generalization of the commonly used Hinge loss that is compatible with both supervised and semi-supervised settings. We study the compromise between training a state of the art generator and an accurate classifier simultaneously, and propose a way to use our algorithm to measure the degree to which a generator and critic are class conditional. We show the trade-off between a generator-critic pair respecting class conditioning inputs and generating the highest quality images. With our multi-hinge loss modification we are able to improve Inception Scores and Frechet Inception Distance on the Imagenet dataset.

****

Driving Among Flatmobiles: Bird-Eye-View Occupancy Grids From a Monocular Camera for Holistic Trajectory Planning

Abdelhak Loukkal, Yves Grandvalet, Tom Drummond, You Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 51-60

Camera-based end-to-end driving neural networks bring the promise of a low-cost system that maps camera images to driving control commands. These networks are appealing because they replace laborious hand engineered building blocks but their black-box nature makes them difficult to delve in case of failure. Recent works have shown the importance of using an explicit intermediate representation that has the benefits of increasing both the interpretability and the accuracy of networks' decisions. Nonetheless, these camera-based networks reason in camera view where scale is not homogeneous and hence not directly suitable for motion forecasting. In this paper, we introduce a novel monocular camera-only holistic end-to-end trajectory planning network with a Bird-Eye-View (BEV) intermediate representation that comes in the form of binary Occupancy Grid Maps (OGMs). To ease the prediction of OGMs in BEV from camera images, we introduce a novel scheme wh ere the OGMs are first predicted as semantic masks in camera view and then warped in BEV using the homography between the two planes. The key element allowing this transformation to be applied to 3D objects such as vehicles, consists in predicting solely their footprint in camera-view, hence respecting the flat world hypothesis implied by the homography.

****

Class-Wise Metric Scaling for Improved Few-Shot Classification

Ge Liu, Linglan Zhao, Wei Li, Dashan Guo, Xiangzhong Fang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 586-595

Few-shot classification aims to generalize basic knowledge to recognize novel categories from a few samples. Recent centroid-based methods achieve promising classification performance with the nearest neighbor rule. However, we consider that those methods intrinsically ignore per-class distribution, as the decision boundaries are biased due to the diversity of intra-class variances. Hence, we propose a class-wise metric scaling (CMS) mechanism, which can be applied to both training and testing stages. Concretely, metric scalars are set as learnable parameters in the training stage, helping to learn a more discriminative and transferable feature representation. As for testing, we construct a convex optimization problem to generate an optimal scalar vector for refining the nearest neighbor decisions. Besides, we also involve a low-ranking bilinear pooling layer for improved representation capacity, which further provides significant performance gains. Extensive experiments are conducted on a series of feature extractor backbones, datasets, and testing modes, which have shown consistent improvements compared to prior SOTA methods, e.g., we achieve accuracies of 66.64% and 83.63% for 5-way 1-shot and 5-shot settings on the mini-ImageNet, respectively. Under the semi-supervised inductive mode, results are further up to 78.34% and 87.53%, respectively.

****

Weakly Supervised Deep Reinforcement Learning for Video Summarization With Semantically Meaningful Reward

Zutong Li, Lei Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3239-3247

Conventional unsupervised video summarization algorithms are usually developed in a frame level clustering manner. For example, frame level diversity and representativeness are two typical clustering criteria used for unsupervised reinforcement learning-based video summarization. Inspired by recent progress in video representation techniques, we further introduce the similarity of video representations to construct a semantically meaningful reward for this task. We consider that a good summarization should also be semantically identical to its original source, which means that the semantic similarity can be regarded as an additional criterion for summarization. Through combining a novel video semantic reward with other unsupervised rewards for training, we can easily upgrade an unsupervised reinforcement learning-based video summarization method to its weakly supervised version. In practice, we first train a video classification sub-network (VCSN) to extract video semantic representations based on a category-labeled video dataset. Then we fix this VCSN and train a summary generation sub-network (SGSN) using unlabeled video data in a reinforcement learning way. Experimental results demonstrate that our work significantly surpasses other unsupervised and even supervised methods. To the best of our knowledge, our method achieves state-of-the-art performance in terms of the correlation coefficients, Kendall's \tau and Spearman's \rho.

****

InfoMax-GAN: Improved Adversarial Image Generation via Information Maximization and Contrastive Learning

Kwot Sin Lee, Ngoc-Trung Tran, Ngai-Man Cheung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3942-3952

While Generative Adversarial Networks (GANs) are fundamental to many generative modelling applications, they suffer from numerous issues. In this work, we propose a principled framework to simultaneously mitigate two fundamental issues in G

ANs: catastrophic forgetting of the discriminator and mode collapse of the generator. We achieve this by employing for GANs a contrastive learning and mutual information maximization approach, and perform extensive analyses to understand sources of improvements. Our approach significantly stabilizes GAN training and improves GAN performance for image synthesis across five datasets under the same training and evaluation conditions against state-of-the-art works. In particular, compared to the state-of-the-art SSGAN, our approach does not suffer from poorer performance on image domains such as faces, and instead improves performance significantly. Our approach is simple to implement and practical: it involves only one auxiliary objective, has low computational cost, and performs robustly across a wide range of training settings and datasets without any hyperparameter tuning. For reproducibility, our code is available in the open-source GAN library, Mimicry.

****

Can Selfless Learning Improve Accuracy of a Single Classification Task?

Soumya Roy, Bharat Bhusan Sau; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4044-4052

The human brain has billions of neurons. However, we perform tasks using only a few concurrently active neurons. Moreover, an activated neuron inhibits the activity of its neighbors. Selfless Learning exploits these neurobiological principles to solve the problem of catastrophic forgetting in continual learning. In this paper, we ask a basic question: can the selfless learning idea be used to improve the accuracy of deep convolutional networks on a single classification task? To achieve this goal, we introduce two regularizers and formulate a curriculum learning-esque strategy to effectively enforce these regularizers on a network. This has resulted in significant gains over vanilla cross-entropy training. Moreover, we have shown that our method can be used in conjunction with other popular learning paradigms like curriculum learning.

****

Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression

Souvik Kundu, Gourav Datta, Massoud Pedram, Peter A. Beerel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3953-3962

The increasing demand for on-chip edge intelligence has motivated the exploration of algorithmic techniques and specialized hardware to reduce the computing energy of current machine learning models. In particular, deep spiking neural networks (SNNs) have gained interest because their event-driven hardware implementations can consume very low energy. However, minimizing average spiking activity and thus energy consumption while preserving accuracy in deep SNNs remains a significant challenge and opportunity. This paper proposes a novel two-step SNN compression technique to reduce their spiking activity while maintaining accuracy that involves compressing specifically-designed artificial neural networks (ANNs) that are then converted into the target SNNs. Our approach uses an ultra-high ANN compression technique that is guided by the attention-maps of an uncompressed meta-model. We then evaluate the firing threshold of each ANN layer and start with the trained ANN weights to perform a sparse-learning-based supervised SNN training to minimize the number of timesteps required while retaining compression. To evaluate the merits of the proposed approach, we performed experiments with variants of VGG and ResNet, on both CIFAR-10and CIFAR-100, and VGG16 on Tiny-ImageNet. SNN mod-els generated through the proposed technique yield state-of-the-art compression ratios of up to 33.4x with no significant drop in accuracy compared to baseline unpruned counterparts. As opposed to the existing SNN pruning methods we achieve up to 8.3x better compression with no drop inaccuracy. Moreover, compressed SNN models generated by our methods can have up to 12.2x better compute energy-efficiency compared to ANNs that have a similar number of parameters.

****

StacMR: Scene-Text Aware Cross-Modal Retrieval

Andres Mafla, Rafael S. Rezende, Lluis Gomez, Diane Larlus, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi on (WACV), 2021, pp. 2220-2230

Recent models for cross-modal retrieval have benefited from an increasingly rich understanding of visual scenes, afforded by scene graphs and object interactions to mention a few. This has resulted in an improved matching between the visual representation of an image and the textual representation of its caption. Yet, current visual representations overlook a key aspect: the text appearing in images, which may contain crucial information for retrieval. In this paper, we first propose a new dataset that allows exploration of cross-modal retrieval where images contain scene-text instances. Then, armed with this dataset, we describe several approaches which leverage scene text, including a better scene-text aware cross-modal retrieval method which uses specialized representations for text from the captions and text from the visual scene, and reconcile them in a common embedding space. Extensive experiments confirm that cross-modal retrieval approaches benefit from scene text and highlight interesting research questions worth exploring further. Dataset and code are available at europe.naverlabs.com/stacmr.

****

DeepCSR: A 3D Deep Learning Approach for Cortical Surface Reconstruction

Rodrigo Santa Cruz, Leo Lebrat, Pierrick Bourgeat, Clinton Fookes, Jurgen Fripp, Olivier Salvado; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 806-815

The study of neurodegenerative diseases relies on the reconstruction and analysis of the brain cortex from magnetic resonance imaging (MRI). Traditional frameworks for this task like FreeSurfer demand lengthy runtimes, while its accelerated variant FastSurfer still relies on a voxel-wise segmentation which is limited by its resolution to capture narrow continuous objects as cortical surfaces. Having these limitations in mind, we propose DeepCSR, a 3D deep learning framework for cortical surface reconstruction from MRI. Towards this end, we train a neural network model with hypercolumn features to predict implicit surface representations for points in a brain template space. After training, the cortical surface at a desired level of detail is obtained by evaluating surface representations at specific coordinates, and subsequently applying a topology correction algorithm and an isosurface extraction method. Thanks to the continuous nature of this approach and the efficacy of its hypercolumn features scheme, DeepCSR efficiently reconstructs cortical surfaces at high resolution capturing fine details in the cortical folding. Moreover, DeepCSR is as accurate, more precise, and faster than the widely used FreeSurfer toolbox and its deep learning powered variant FastSurfer on reconstructing cortical surfaces from MRI which should facilitate large-scale medical studies and new healthcare applications.

****

Fair Comparison: Quantifying Variance in Results for Fine-Grained Visual Categorization

Matthew Gwilliam, Adam Teuscher, Connor Anderson, Ryan Farrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3309-3318

For the task of image classification, researchers work arduously to develop the next state-of-the-art (SOTA) model, each bench-marking their own performance against that of their predecessors and of their peers. Unfortunately, the metric used most frequently to describe a model's performance, average categorization accuracy, is often used in isolation. As the number of classes increases, such as in fine-grained visual categorization (FGVC), the amount of information conveyed by average accuracy alone dwindles. While its most glaring weakness is its failure to describe the model's performance on a class-by-class basis, average accuracy also fails to describe how performance may vary from one trained model of the same architecture, on the same dataset, to another (both averaged across all categories and at the per-class level). We first demonstrate the magnitude of these variations across models and across class distributions based on attributes of the data, comparing results on different visual domains and different per-class image distributions, including long-tailed distributions and few-shot subsets. We then analyze the impact various FGVC methods have on overall and per-class variance. From this analysis, we both highlight the importance of reporting and co mparing methods based on information beyond overall accuracy, as well as point out techniques that mitigate variance in FGVC results.

****

Subject Guided Eye Image Synthesis With Application to Gaze Redirection

Harsimran Kaur, Roberto Manduchi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 11-20

We propose a method for synthesizing eye images from segmentation masks with a desired style. The style encompasses attributes such as skin color, texture, iris color, and personal identity. Our approach generates an eye image that is consistent with a given segmentation mask and has the attributes of the input style image. We apply our method to data augmentation as well as to gaze redirection. The previous techniques of synthesizing real eye images from synthetic eye images for data augmentation lacked control over the generated attributes. We demonstrate the effectiveness of the proposed method in synthesizing realistic eye images with given characteristics corresponding to the synthetic labels for data augmentation, which is further useful for various tasks such as gaze estimation, eye image segmentation, pupil detection, etc. We also show how our approach can be applied to gaze redirection using only synthetic gaze labels, improving the previous state of the art results. The main contributions of our paper are i) a novel approach for Style-Based eye image generation from segmentation mask; ii) the use of this approach for gaze-redirection without the need for gaze annotated real eye images

****

Hyperrealistic Image Inpainting With Hypergraphs

Gourav Wadhwa, Abhinav Dhall, Subrahmanyam Murala, Usman Tariq; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3912-3921

Image inpainting is a non-trivial task in computer vi-sion due to multiple possibilities for filling the missing data, which may be dependent on the global information of the image. Most of the existing approaches use the attention mechanism to learn the global context of the image. This attention mechanism produces semantically plausible but blurry results because of incapability to capture the global context. In this paper, we introduce hypergraph convolution on spatial features to learn the complex relationship among the data. We introduce a trainable mechanism to connect nodes using hyperedges for hypergraph convolution. To the best of our knowledge, hypergraph convolution have never been used on spatial features for any image-to-image tasks in computer vision. Further, we introduce gated convolution in the discriminator to enforce local consistency in the predicted image. The experiments on Places2, CelebA-HQ, Paris Street View, and Facades datasets, show that our approach achieves state-of-the-art results.

****

Learning Fast Converging, Effective Conditional Generative Adversarial Networks With a Mirrored Auxiliary Classifier

Zi Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2566-2575

Training conditional generative adversarial networks (GANs) has been remaining as a challenging task, though standard GANs have developed substantially and gained huge successes in recent years. In this paper, we propose a novel conditional GAN architecture with a mirrored auxiliary classifier (MAC-GAN) in its discriminator for the purpose of label conditioning. Unlike existing works, our mirrored auxiliary classifier contains both a real and a fake node for each specific class to distinguish real samples from generated samples that are assigned into the same category by previous models. Comparing with previous auxiliary classifier-based conditional GANs, our MAC-GAN learns a fast converging model for high-quality image generation, taking benefits from its robust, newly designed auxiliary classifier. Experiments on multiple benchmark datasets illustrate that our proposed model improves the quality of image synthesis compared with state-of-the-art approaches. Moreover, much better classification performance can be achieved with the mirrored auxiliary classifier, which can in turn promote the use of MAC-GAN in various transfer learning tasks.

****

Class-Agnostic Object Detection

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, Premkumar Natarajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 919-928

Object detection models perform well at localizing and classifying objects that they are shown during training. However, due to the difficulty and cost associated with creating and annotating detection datasets, trained models detect a limited number of object types with unknown objects treated as background content. This hinders the adoption of conventional detectors in real-world applications like large-scale object matching, visual grounding, visual relation prediction, obstacle detection (where it is more important to determine the presence and location of objects than to find specific types), etc. We propose class-agnostic object detection as a new problem that focuses on detecting objects irrespective of their object-classes. Specifically, the goal is to predict bounding boxes for all objects in an image but not their object-classes. The predicted boxes can then be consumed by another system to perform application-specific classification, retrieval, etc. We propose training and evaluation protocols for benchmarking class-agnostic detectors to advance future research in this domain. Finally, we propose (1) baseline methods and (2) a new adversarial learning framework for class-agnostic detection that forces the model to exclude class-specific information from features used for predictions. Experimental results show that adversarial learning improves class-agnostic detection efficacy.

****

Self-Distillation for Few-Shot Image Captioning

Xianyu Chen, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 545-555

The development of large-scale image-captioning datasets is expensive, while the abundance of unpaired images and text corpus can potentially help reduce the efforts of manual annotation. In this paper, we study the few-shot image captioning problem that only requires a small amount of annotated image-caption pairs. We propose an ensemble-based self-distillation method that allows image captioning models to be trained with unpaired images and captions. The ensemble consists of multiple base models trained with different data samples in each iteration. For learning from unpaired images, we generate multiple pseudo captions with the ensemble and allocate different weights according to their confidence levels. For learning from unpaired captions, we propose a simple yet effective pseudo feature generation method based on Gradient Descent. The pseudo captions and pseudo features from the ensemble are used to train the base models in future iterations. The proposed method is general over different image captioning models and datasets. Our experiments demonstrate significant performance improvements and meaningful captions generated with only 1% of paired training data. Source code is available at https://github.com/chenxy99/SD-FSIC.

****

CenterFusion: Center-Based Radar and Camera Fusion for 3D Object Detection

Ramin Nabati, Hairong Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1527-1536

The perception system in autonomous vehicles is respon-sible for detecting and tracking the surrounding objects.This is usually done by taking advantage of several sens-ing modalities to increase robustness and accuracy, whichmakes sensor fusion a crucial part of the perception system.In this paper, we focus on the problem of radar and cam-era sensor fusion and propose a middle-fusion approachto exploit both radar and camera data for 3D object de-tection. Our approach, called CenterFusion, first uses acenter point detection network to detect objects by identifying their center points on the image. It then solves the key data association problem using a novel frustum-based method to associate the radar detections to their corresponding object's center point. The associated radar detections are used to generate radar-based feature maps to complement the image features, and regress to object properties such as depth, rotation and velocity. We evaluateCenterFusion on the challenging nuScenes dataset, where it improves the overal l nuScenes Detection Score (NDS) of the state-of-the-art camera-based algorithm by more than12%. We further show that CenterFusion significantly improves the velocity estimation accuracy without using any additional temporal information. The code is available at https://github.com/mrnabati/CenterFusion.

****

Viewpoint-Agnostic Image Rendering

Hiroaki Aizawa, Hirokatsu Kataoka, Yutaka Satoh, Kunihito Kato; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3803-3812

Rendering an any-viewpoint image is extremely difficult for Generative Adversarial Networks. This is because conventional GANs do not understand 3D information underlying a given viewpoint image such as an object shape and relationship between viewpoint and objects in 3D space. In this paper, we present how to perform a Viewpoint-Agnostic Image Rendering (VAIR), equipping a conditional GAN with a mechanism to reconstruct 3D information of the input view. VAIR realizes any-viewpoint image generation by manipulating a viewpoint in 3D space where the reconstructed instance shape is arranged. In addition, we convert the reconstructed 3D shape into a 2D representation for image-based conditional GAN, while preserving detail 3D information. The representation consists of a depth image and 2D semantic keypoint images, which are obtained by rendering the shape from a viewpoint. In the experiment, we evaluate using a CUB-200-2011 dataset, which contains few-samples biased a viewpoint such that covers only part of the target appearance. As a result, our VAIR clearly renders an any-viewpoint image.

****

EAGLE-Eye: Extreme-Pose Action Grader Using Detail Bird's-Eye View

Mahdiar Nekoui, Fidel Omar Tito Cruz, Li Cheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 394-402

Measuring the quality of a sports action entails attending to the execution of the short-term components as well as the overall impression of the whole program. In this assessment, both appearance clues and pose dynamics features should be involved. Current approaches often treat a sports routine as a simple fine-grained action, while taking little heed of its complex temporal structure. Besides, most of them rely solely on either appearance or pose features to score the performance. In this paper, we present JCA and ADA blocks that are responsible for reasoning about the coordination among the joints and appearance dynamics throughout the performance. We build our two-stream network upon the separate stack of these blocks. The early blocks capture the fine-grained temporal dependencies while the last ones reason about the long-term coarse-grained relations. We further introduce an annotated dataset of sports images with unusual pose configurations to boost the performance of pose estimation in such scenarios. Our experiments show that the proposed method not only outperforms the previous works in short - term action assessment but also is the first to generalize well to minute-long figure-skating scoring.

****

CAP: Context-Aware Pruning for Semantic Segmentation

Wei He, Meiqing Wu, Mingfu Liang, Siew-Kei Lam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 960-969

Network pruning for deep convolutional neural networks (CNNs) has recently achieved notable research progress on image-level classification. However, most existing pruning methods are not catered to or evaluated on semantic segmentation networks. In this paper, we advocate the importance of contextual information during channel pruning for semantic segmentation networks by presenting a novel Context-aware Pruning framework. Concretely, we formulate the embedded contextual information by leveraging the layer-wise channels interdependency via the Context-aware Guiding Module (CAGM) and introduce the Context-aware Guided Sparsification (CAGS) to adaptively identify the informative channels on the cumbersome model by inducing channel-wise sparsity on the scaling factors in batch normalization (BN) layers. The resulting pruned models require significantly lesser operations for inference while maintaining comparable performance to (at times outperforming) the original models. We evaluated our framework on widely-used benchmarks an d showed its effectiveness on both large and lightweight models. On Cityscapes dataset, our framework reduces the number of parameters by 32%, 47%, 54%, and 63%, on PSPNet101, PSPNet50, ICNet, and SegNet, respectively, while preserving the performance.

****

Breaking Shortcuts by Masking for Robust Visual Reasoning

Keren Ye, Mingda Zhang, Adriana Kovashka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3520-3530

Visual reasoning is a challenging but important task that is gaining momentum. Examples include reasoning about what will happen next in film, or interpreting what actions an image advertisement prompts. Both tasks are "puzzles" which invite the viewer to combine knowledge from prior experience, to find the answer. Intuitively, providing external knowledge to a model should be helpful, but it does not necessarily result in improved reasoning ability. An algorithm can learn to find answers to the prediction task yet not perform generalizable reasoning. In other words, models can leverage "shortcuts" between inputs and desired outputs, to bypass the need for reasoning. We develop a technique to effectively incorporate external knowledge, in a way that is both interpretable, and boosts the contribution of external knowledge for multiple complementary metrics. In particular, we mask evidence in the image and in retrieved external knowledge. We show this masking successfully focuses the method's attention on patterns that generalize. To properly understand how our method utilizes external knowledge, we propose a novel side evaluation task. We find that with our masking technique, the model can learn to select useful knowledge pieces to rely on.

****

Domain-Aware Unsupervised Hyperspectral Reconstruction for Aerial Image Dehazing

Aditya Mehta, Harsh Sinha, Murari Mandal, Pratik Narang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 413-422

Haze removal in aerial images is a challenging problem due to considerable variation in spatial details and varying contrast. Changes in particulate matter density often lead to degradation in visibility. Therefore, several approaches utilize multi-spectral data as auxiliary information for haze removal. In this paper, we propose SkyGAN for haze removal in aerial images. SkyGAN comprises of 1) a domain-aware hazy-to-hyperspectral (H2H) module, and 2) a conditional GAN (cGAN) based multi-cue image-to-image translation module (I2I) for dehazing. The proposed H2H module reconstructs several visual bands from RGB images in an unsupervised manner, which overcomes the lack of hazy hyperspectral aerial image datasets. The module utilizes task supervision and domain adaptation in order to create a "hyperspectral catalyst" for image dehazing. The I2I module uses the hyperspectral catalyst along with a 12-channel multi-cue input and performs effective image dehazing by utilizing the entire visual spectrum. In addition, this work introduces a new dataset, called Hazy Aerial-Image (HAI) dataset, that contains more than 65,000 pairs of hazy and ground truth aerial images with realistic, non-homogeneous haze of varying density. The performance of SkyGAN is evaluated on the recent SateHaze1k dataset as well as the HAI dataset. We also present a comprehensive evaluation of HAI dataset with a representative set of state-of-the-art techniques in terms of PSNR and SSIM.

****

Cross-Domain Latent Modulation for Variational Transfer Learning

Jinyong Hou, Jeremiah D. Deng, Stephen Cranefield, Xuejie Ding; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3149-3158

We propose a cross-domain latent modulation mechanism within a variational autoencoders (VAE) framework to enable improved transfer learning. Our key idea is to procure deep representations from one data domain and use it as perturbation to the reparameterization of the latent variable in another domain. Specifically, deep representations of the source and target domains are first extracted by a unified inference model and aligned by employing gradient reversal. Second, the learned deep representations are cross-modulated to the latent encoding of the al ternate domain. The consistency between the reconstruction from the modulated latent encoding and the generation using deep representation samples is then enforced in order to produce inter-class alignment in the latent space further. We apply the proposed model to a number of transfer learning tasks including unsupervised domain adaptation and image-to-image translation. Experimental results show that our model gives competitive performance.

****

DeepOpht: Medical Report Generation for Retinal Images via Deep Models and Visual Explanation

Jia-Hong Huang, C.-H. Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I-Hung Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, Jesper Tegner, Marcel Worring; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2442-2452

In this work, we propose an AI-based method that intends to improve the conventional retinal disease treatment procedure and help ophthalmologists increase diagnosis efficiency and accuracy. The proposed method is composed of a deep neural networks-based (DNN-based) module, including a retinal disease identifier and clinical description generator, and a DNN visual explanation module. To train and validate the effectiveness of our DNN-based module, we propose a large-scale retinal disease image dataset. Also, as ground truth, we provide a retinal image dataset manually labeled by ophthalmologists to qualitatively show the proposed AI-based method is effective. With our experimental results, we show that the proposed method is quantitatively and qualitatively effective. Our method is capable of creating meaningful retinal image descriptions and visual explanations that are clinically relevant.

****

Generating Physically Sound Training Data for Image Recognition of Additively Manufactured Parts

Tobias Nickchen, Stefan Heindorf, Gregor Engels; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1994-2002

In recent years, Additive Manufacturing (AM) has evolved from a niche technology for prototyping to a well-known industrial production process. In this work, we focus on Selective Laser Sintering (SLS)---one of the leading AM techniques. While SLS has many advantages, the simultaneous manufacturing of multiple components requires the subsequent recognition of components which must be done manually in today's production processes. While approaches for automatic, sensor-based object recognition have been proposed, e.g., based on Convolutional Neural Networks (CNNs), they assume the availability of real-world photos which is not given in the setting of Additive Manufacturing. Hence, we develop an approach to render realistic virtual images and demonstrate their suitability to recognize real-world objects. Although often done in the machine learning community, orienting the objects randomly generates many orientations that are physically impossible and cause distracting noise in the training process. Hence, we pay particular attention to generate physically sound training data and we demonstrate that our approach significantly improves the recognition rate compared to traditional approaches.

****

GraphTCN: Spatio-Temporal Interaction Modeling for Human Trajectory Prediction

Chengxin Wang, Shaofeng Cai, Gary Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3450-3459

Predicting the future paths of an agent's neighbors accurately and in a timely manner is central to the autonomous applications for collision avoidance. Conventional approaches, e.g., LSTM-based models, take considerable computational costs in the prediction, especially for the long sequence prediction. To support more efficient and accurate trajectory predictions, we propose a novel CNN-based spatial-temporal graph framework GraphTCN, which models the spatial interactions as social graphs and captures the spatio-temporal interactions with a modified temporal convolutional network. In contrast to conventional models, both the spatial and temporal modeling of our model are computed within each local time window. Therefore, it can be executed in parallel for much higher efficiency, and meanw hile with accuracy comparable to best-performing approaches. Experimental results confirm that our model achieves better performance in terms of both efficiency and accuracy as compared with state-of-the-art models on various trajectory prediction benchmark datasets.

****

Hand Pose Guided 3D Pooling for Word-Level Sign Language Recognition

Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, Jana Kosecka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3429-3439

Gestures in American Sign Language (ASL) are characterized by fast, highly articulate motion of upper body, including arm movements with complex hand shapes and facial expressions. In this work, we propose a new method for word-level sign recognition from American Sign Language (ASL) using video. Our method uses both motion and hand shape cues while being robust to variations of execution. We exploit the knowledge of the body pose, estimated from an off-the-shelf pose estimator. Using the pose as a guide, we pool spatio-temporal feature maps from different layers of a 3D convolutional neural network. We train separate classifiers using pose guided pooled features from different resolutions and fuse their prediction scores during test time. This leads to a significant improvement in performance on the WLASL benchmark dataset [25]. The proposed approach achieves 10%, 12%, 9:5% and 6:5% performance gain on WLASL100, WLASL300, WLASL1000, WLASL2000 subsets respectively. To demonstrate the robustness of the pose guided pooling and proposed fusion mechanism, we also evaluate our method by fine tuning the model on another dataset. This yields 10% performance improvement for the proposed method using only 0:4% training data during fine tuning stage.

****

Efficient Attention: Attention With Linear Complexities

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, Hongsheng Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3531-3539

Dot-product attention has wide applications in computer vision and natural language processing. However, its memory and computational costs grow quadratically with the input size. Such growth prohibits its application on high-resolution inputs. To remedy this drawback, this paper proposes a novel efficient attention mechanism equivalent to dot-product attention but with substantially less memory and computational costs. Its resource efficiency allows more widespread and flexible integration of attention modules into a network, which leads to better accuracies. Empirical evaluations demonstrated the effectiveness of its advantages. Models with efficient attention achieved state-of-the-art accuracies on MS-COCO 2017. Further, the resource efficiency democratizes attention to complex models, where high costs prohibit the use of dot-product attention. As an exemplar, a model with efficient attention achieved state-of-the-art accuracies for stereo depth estimation on the Scene Flow dataset. Code is available at https://github.com/cmsflash/efficient-attention.

****

Enhancing Diversity in Teacher-Student Networks via Asymmetric Branches for Unsupervised Person Re-Identification

Hao Chen, Benoit Lagadec, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1-10

The objective of unsupervised person re-identification (Re-ID) is to learn discriminative features without labor-intensive identity annotations. State-of-the-art unsupervised Re-ID methods assign pseudo labels to unlabeled images in the target domain and learn from these noisy pseudo labels. Recently introduced Mean Teacher Model is a promising way to mitigate the label noise. However, during the training, self-ensembled teacher-student networks quickly converge to a consensus which leads to a local minimum. We explore the possibility of using an asymmetric structure inside neural network to address this problem. First, asymmetric branches are proposed to extract features in different manners, which enhances the feature diversity in appearance signatures. Then, our proposed cross-branch supervision allows one branch to get supervision from the other branch, which tran sfers distinct knowledge and enhances the weight diversity between teacher and student networks. Extensive experiments show that our proposed method can significantly surpass the performance of previous work on both unsupervised domain adaptation and fully unsupervised Re-ID tasks.

****

Large Image Datasets: A Pyrrhic Win for Computer Vision?
Abeba Birhane, Vinay Uday Prabhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1537-1547

In this paper we investigate problematic practices and consequences of large scale vision datasets (LSVDs). We examine broad issues such as the question of consent and justice as well as specific concerns such as the inclusion of verifiably pornographic images in datasets. Taking the ImageNet-ILSVRC-2012 dataset as an example, we perform a cross-sectional model-based quantitative census covering factors such as age, gender, NSFW content scoring, class-wise accuracy, human-cardinality-analysis, and the semanticity of the image class information in order to statistically investigate the extent and subtleties of ethical transgressions. We then use the census to help hand-curate a look-up-table of images in the ImageNet-ILSVRC-2012 dataset that fall into the categories of verifiably pornographic: shot in a non-consensual setting (up-skirt), beach voyeuristic, and exposed private parts. We survey the landscape of harm and threats both the society at large and individuals face due to uncritical and ill-considered dataset curation practices. We then propose possible courses of correction and critique their pros and cons. We have duly open-sourced all of the code and the census meta-datasets generated in this endeavor for the computer vision community to build on.

****

Triangle-Net: Towards Robustness in Point Cloud Learning
Chenxi Xiao, Juan Wachs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 826-835

Three dimensional (3D) object recognition is becoming a key desired capability for many computer vision systems such as autonomous vehicles, service robots and surveillance drones to operate more effectively in unstructured environments. These real-time systems require effective classification methods that are robust to various sampling resolutions, noisy measurements, and unconstrained pose configurations. Previous research has shown that points' sparsity, rotation and positional inherent variance can lead to a significant drop in the performance of point cloud based classification techniques. However, neither of them is sufficiently robust to multifactorial variance and significant sparsity. In this regard, we propose a novel approach for 3D classification that can simultaneously achieve invariance towards rotation, positional shift, scaling, and is robust to point sparsity. To this end, we introduce a new feature that utilizes graph structure of point clouds, which can be learned end-to-end with our proposed neural network to acquire a robust latent representation of the 3D object. We show that such latent representations can significantly improve the performance of object classification and retrieval tasks when points are sparse. Further, we show that our approach outperforms PointNet and 3DmFV by 35.0% and 28.1% respectively in ModelNet 40 classification tasks using sparse point clouds of only 16 points under arbitrary SO(3) rotation.

****

LT-GAN: Self-Supervised GAN With Latent Transformation Detection
Parth Patel, Nupur Kumari, Mayank Singh, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3189-3198

Generative Adversarial Networks (GANs) coupled with self-supervised tasks have shown promising results in unconditional and semi-supervised image generation. We propose a self-supervised approach (LT-GAN) to improve the generation quality and diversity of images by estimating the GAN-induced transformation (i.e. transformation induced in the generated images by perturbing the latent space of generator). Specifically, given two pairs of images where each pair comprises of a generated image and its transformed version, the self-supervision task aims to identify whether the latent transformation applied in the given pair is same as tha t of the other pair. Hence, this auxiliary loss encourages the generator to produce images that are distinguishable by the auxiliary network, which in turn promotes the synthesis of semantically consistent images with respect to latent transformations. We show the efficacy of this pretext task by improving the image generation quality in terms of FID on state-of-the-art models in conditional and unconditional settings on CIFAR-10, CelebA-HQ and ImageNet datasets. Moreover, we empirically show that LT-GAN helps in improving controlled image editing for CelebA-HQ, and ImageNet over baseline models. We experimentally demonstrate that our proposed LT self-supervision task can be effectively combined with other state-of-the-art training techniques for added benefits. Consequently, we show that our approach achieves the new state-of-the-art FID score of 9.8 on conditional CIFAR-10 image generation.

****

Audio- and Gaze-Driven Facial Animation of Codec Avatars

Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, Yaser Sheikh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 41-50

Codec Avatars are a recent class of learned, photorealistic face models that accurately represent the geometry and texture of a person in 3D (i.e., for virtual reality), and are almost indistinguishable from video. In this paper we describe the first approach to animate these parametric models in real-time which could be deployed on commodity virtual reality hardware using audio and/or eye tracking. Our goal is to display expressive conversations between individuals that exhibit important social signals such as laughter and excitement solely from latent cues in our lossy input signals. To this end we collected over 5 hours of high frame rate 3D face scans across three participants including traditional neutral speech as well as expressive and conversational speech. We investigate a multimodal fusion approach that dynamically identifies which sensor encoding should animate which parts of the face at any time. See the supplemental video which demonstrates our ability to generate full face motion far beyond the typically neutral lip articulations seen in competing work: https://research.fb.com/videos/audio-and-gaze-driven-facial-animation-of-codec-avatars/

****

Transductive Zero-Shot Learning by Decoupled Feature Generation

Federico Marmoreo, Jacopo Cavazza, Vittorio Murino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3109-3118

In this paper, we address zero-shot learning (ZSL), the problem of recognizing categories for which no labeled visual data are available during training. We focus on the transductive setting, in which unlabelled visual data from unseen classes is available. State-of-the-art paradigms in ZSL typically exploit generative adversarial networks to synthesize visual features from semantic attributes. We posit that the main limitation of these approaches is to adopt a single model to face two problems: 1) generating realistic visual features, and 2) translating semantic attributes into visual cues. Differently, we propose to decouple such tasks, solving them separately. In particular, we train an unconditional generator to solely capture the complexity of the distribution of visual data and we subsequently pair it with a conditional generator devoted to enrich the prior knowledge of the data distribution with the semantic content of the class embeddings. We present a detailed ablation study to dissect the effect of our proposed decoupling approach, while demonstrating its superiority over the related state-of-the-art.

****

Two-Hand Global 3D Pose Estimation Using Monocular RGB

Fanqing Lin, Connor Wilhelm, Tony Martinez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2373-2381

We tackle the challenging task of estimating global 3D joint locations for both hands via only monocular RGB input images. We propose a novel multi-stage convolutional neural network based pipeline that accurately segments and locates the hands despite occlusion between two hands and complex background noise and estimates the 2D and 3D canonical joint locations without any depth information. Globa

l joint locations with respect to the camera origin are computed using the hand pose estimations and the actual length of the key bone with a novel projection algorithm. To train the CNNs for this new task, we introduce a large-scale synthetic 3D hand pose dataset. We demonstrate that our system outperforms previous works on 3D canonical hand pose estimation benchmark datasets with RGB-only information. Additionally, we present the first work that achieves accurate global 3D hand tracking on both hands using RGB-only inputs and provide extensive quantitative and qualitative evaluation.
****

The Devil Is in the Boundary: Exploiting Boundary Representation for Basis-Based Instance Segmentation

Myungchul Kim, Sanghyun Woo, Dahun Kim, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 929-938

Pursuing a more coherent scene understanding towards real-time vision applications, single-stage instance segmentation has recently gained popularity, achieving a simpler and more efficient design than its two-stage counterparts. Besides, its global mask representation often leads to superior accuracy to the two-stage Mask R-CNN which has been dominant thus far. Despite the promising advances in single-stage methods, finer delineation of instance boundaries still remains unexcavated. Indeed, boundary information provides a strong shape representation that can operate in synergy with the fully-convolutional mask features of the single-stage segmented. In this work, we propose Boundary Basis based Instance Segmentation(B2Inst) to learn a global boundary representation that can complement existing global-mask-based methods that are often lacking high-frequency details. Besides, we devise a unified quality measure of both mask and boundary and introduce a network block that learns to score the per-instance predictions of itself. When applied to the strongest baselines in single-stage instance segmentation, our B2Inst leads to consistent improvements and accurately parse out the instance boundaries in a scene. Regardless of being single-stage or two-stage frameworks, we outperform the existing state-of-the-art methods on the COCO dataset with the same ResNet-50 and ResNet-101 backbones.
****

Towards Zero-Shot Learning With Fewer Seen Class Examples

Vinay Kumar Verma, Ashish Mishra, Anubha Pandey, Hema A. Murthy, Piyush Rai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2241-2251

We present a meta-learning based generative model for zero-shot learning (ZSL) towards a challenging setting when the number of training examples from each seen class is very few. This setup is in contrast to the conventional ZSL approaches, where training typically assumes the availability of a sufficiently large number of training examples from each of the seen classes. The proposed approach leverages meta-learning to train a deep generative model that integrates variational autoencoder an generative adversarial network. To simulate the ZSL behaviour in training, we propose a novel task distribution where meta-train and meta-validation classes are disjoint. Once trained, the model can generate synthetic examples from seen and unseen classes. Synthesize samples can then be used to train the ZSL framework in a supervised manner. The meta-learner enables our model to generates high-fidelity samples using only a small number of training examples from seen classes. We conduct extensive experiments and ablation studies on four benchmark datasets of ZSL and observe that the proposed model outperforms state-of-the-art approaches by a significant margin when the number of examples per seen class is very small.
****

Painting Outside As Inside: Edge Guided Image Outpainting via Bidirectional Rearrangement With Progressive Step Learning

Kyunghun Kim, Yeohun Yun, Keon-Woo Kang, Kyeongbo Kong, Siyeong Lee, Suk-Ju Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2122-2130

Image outpainting is a very intriguing problem as the outside of a given image can be continuously filled by considering as the context of the image. This task has two main challenges. The first is to maintain the spatial consistency in contents of generated regions and the original input. The second is to generate a high-quality large image with a small amount of adjacent information. Conventional image outpainting methods generate inconsistent, blurry, and repeated pixels. To alleviate the difficulty of an outpainting problem, we propose a novel image outpainting method using bidirectional boundary region rearrangement. We rearrange the image to benefit from the image inpainting task by reflecting more directional information. The bidirectional boundary region rearrangement enables the generation of the missing region using bidirectional information similar to that of the image inpainting task, thereby generating the higher quality than the conventional methods using unidirectional information. Moreover, we use the edge map generator that considers images as original input with structural information and hallucinates the edges of unknown regions to generate the image. Our proposed method is compared with other state-of-the-art outpainting and inpainting methods both qualitatively and quantitatively. We further compared and evaluated them using BRISQUE, one of the No-Reference image quality assessment (IQA) metrics, to evaluate the naturalness of the output. The experimental results demonstrate that our method outperforms other methods and generates new images with 360degpanoramic characteristics.

****

End-to-End Learning Improves Static Object Geo-Localization From Video

Mohamed Chaabane, Lionel Gueguen, Ameni Trabelsi, Ross Beveridge, Stephen O'Hara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2063-2072

Accurately estimating the position of static objects, such as traffic lights, from the moving camera of a self-driving car is a challenging problem. In this work, we present a system that improves the localization of static objects by jointly-optimizing the components of the system via learning. Our system is comprised of networks that perform: 1) 5DoF object pose estimation from a single image, 2) association of objects between pairs of frames, and 3) multi-object tracking to produce the final geo-localization of the static objects within the scene. We evaluate our approach using a publicly-available data set, focusing on traffic lights due to data availability. For each component, we compare against contemporary alternatives and show significantly-improved performance. We also show that the end-to-end system performance is further improved via joint-training of the constituent models.

****

Self-Supervised 4D Spatio-Temporal Feature Learning via Order Prediction of Sequential Point Cloud Clips

Haiyan Wang, Liang Yang, Xuejian Rong, Jinglun Feng, Yingli Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3762-3771

Recently 3D scene understanding attracts attention for many applications, however, annotating a vast amount of 3D data for training is usually expensive and time-consuming. To alleviate the needs of ground truth, we propose a self-supervised schema to learn 4D spatio-temporal features (i.e. 3 spatial dimensions plus 1 temporal dimension) from dynamic point cloud data by predicting the temporal order of sampled and shuffled point cloud clips. 3D sequential point cloud contains precious geometric and depth information to better recognize activities in 3D space compared to videos. To learn the 4D spatio-temporal features, we introduce 4D convolution neural networks to predict the temporal order on a self-created large scale dataset, NTU-PCLs, derived from the NTU-RGB+D dataset. The efficacy of the learned 4D spatio-temporal features is verified on two tasks: 1) Self-supervised 3D nearest neighbor retrieval; and 2) Self-supervised representation learning transferred for action recognition on the smaller 3D dataset. Our extensive experiments prove the effectiveness of the proposed self-supervised learning method which achieves comparable results w.r.t. the fully-supervised methods on action recognition on MSRAction3D dataset.

****

Domain Impression: A Source Data Free Domain Adaptation Method

Vinod K. Kurmi, Venkatesh K. Subramanian, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 615-625

Unsupervised Domain adaptation methods solve the adaptation problem for an unlabeled target set, assuming that the source dataset is available with all labels. However, the availability of actual source samples is not always possible in practical cases. It could be due to memory constraints, privacy concerns, and challenges in sharing data. This practical scenario creates a bottleneck in the domain adaptation problem. This paper addresses this challenging scenario by proposing a domain adaptation technique that does not need any source data. Instead of the source data, we are only provided with a classifier that is trained on the source data. Our proposed approach is based on a generative framework, where the trained classifier is used for generating samples from the source classes. We learn the joint distribution of data by using the energy-based modeling of the trained classifier. At the same time, a new classifier is also adapted for the target domain. We perform various ablation analysis under different experimental setups and demonstrate that the proposed approach achieves better results than the baseline models in this extremely novel scenario.

****

Identity Unbiased Deception Detection by 2D-to-3D Face Reconstruction

Le Minh Ngo, Wei Wang, Burak Mandira, Sezer Karaoglu, Henri Bouma, Hamdi Dibeklioglu, Theo Gevers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 145-154

Deception is a common phenomenon in society, both in our private and professional lives. However, humans are notoriously bad at accurate deception detection. Based on the literature, human accuracy of distinguishing between lies and truthful statements is 54% on average, in other words, it is slightly better than a random guess. While people do not much care about this issue, in high-stakes situations such as interrogations for series crimes and for evaluating the testimonies in court cases, accurate deception detection methods are highly desirable. To achieve a reliable, covert, and non-invasive deception detection, we propose a novel method that disentangles facial expression and head pose related features using 2D-to-3D face reconstruction technique from a video sequence and uses them to learn characteristics of deceptive behavior. We evaluate the proposed method on the Real-Life Trial (RLT) dataset that contains high-stakes deceits recorded in courtrooms. Our results show that the proposed method (with an accuracy of 68%) improves the state of the art. Besides, a new dataset has been collected, for the first time, for low-stake deceit detection. In addition, we compare high-stake deceit detection methods on the newly collected low-stake deceits.

****

Phase-Wise Parameter Aggregation for Improving SGD Optimization

Takumi Kobayashi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2625-2634

Stochastic gradient descent (SGD) is successfully applied to train deep convolutional neural networks (CNNs) on various computer vision tasks. Since fixed step-size SGD converges to so-called error plateau, it is applied in combination with decaying learning rate to reach a favorable optimum. In this paper, we propose a simple yet effective optimization method to improve SGD with a phase-wise decay of learning rate. Through analyzing both a loss surface around the error plateau and a structure of the SGD optimization process, the proposed method is formulated to improve convergence as well as initialization at each training phase by efficiently aggregating the CNN parameters along the optimization sequence. The method keeps the simplicity of SGD while touching the SGD procedure only a few times during training. The experimental results on image classification tasks thoroughly validate the effectiveness of the proposed method in comparison to the other methods.

****

Learn Like a Pathologist: Curriculum Learning by Annotator Agreement for Histopathology Image Classification

Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Va ickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, Lorenzo Torresani, Jason Wei, Saeed Hassanpour; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2473-2483

Applying curriculum learning requires both a range of difficulty in data and a method for determining the difficulty of examples. In many tasks, however, satisfying these requirements can be a formidable challenge. In this paper, we contend that histopathology image classification is a compelling use case for curriculum learning. Based on the nature of histopathology images, a range of difficulty inherently exists among examples, and, since medical datasets are often labeled by multiple annotators, annotator agreement can be used as a natural proxy for the difficulty of a given example. Hence, we propose a simple curriculum learning method that trains on progressively-harder images as determined by annotator agreement. We evaluate our hypothesis on the challenging and clinically-important task of colorectal polyp classification. Whereas vanilla training achieves an AUC of 83.7% for this task, a model trained with our proposed curriculum learning approach achieves an AUC of 88.2%, an improvement of 4.5%. Our work aims to inspire researchers to think more creatively and rigorously when choosing contexts for applying curriculum learning.

****

Illumination Normalization by Partially Impossible Encoder-Decoder Cost Function

Steve Dias Da Cruz, Bertram Taetz, Thomas Stifter, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1459-1468

Images recorded during the lifetime of computer vision based systems undergo a wide range of illumination and environmental conditions affecting the reliability of previously trained machine learning models. Image normalization is hence a valuable preprocessing component to enhance the models' robustness. To this end, we introduce a new strategy for the cost function formulation of encoder-decoder networks to average out all the unimportant information in the input images (e.g. environmental features and illumination changes) to focus on the reconstruction of the salient features (e.g. class instances). Our method exploits the availability of identical sceneries under different illumination and environmental conditions for which we formulate a partially impossible reconstruction target: the input image will not convey enough information to reconstruct the target in its entirety. Its applicability is assessed on three publicly available datasets. We combine the triplet loss as a regularizer in the latent space representation and a nearest neighbour search to improve the generalization to unseen illuminations and class instances. The importance of the aforementioned post-processing is highlighted on an automotive application. To this end, we release a synthetic dataset of sceneries from three different passenger compartments where each scenery is rendered under ten different illumination and environmental conditions: https://sviro.kl.dfki.de

****

RarePlanes: Synthetic Data Takes Flight

Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, Daeil Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 207-217

RarePlanes is a unique open-source machine learning dataset that incorporates both real and synthetically generated satellite imagery. The RarePlanes dataset specifically focuses on the value of synthetic data to aid computer vision algorithms in their ability to automatically detect aircraft and their attributes in satellite imagery. Although other synthetic/real combination datasets exist, RarePlanes is the largest openly-available very-high resolution dataset built to test the value of synthetic data from an overhead perspective. Previous research has shown that synthetic data can reduce the amount of real training data needed and potentially improve performance for many tasks in the computer vision domain. The real portion of the dataset consists of 253 Maxar WorldView-3 satellite scenes spanning 112 locations and 2,142 km^2 with 14,700 hand-annotated aircraft. The accompanying synthetic dataset is generated via AI.Reverie's simulation platform and features 50,000 synthetic satellite images simulating a total area of 933

1.2 km^2 with 630,000 aircraft annotations. Both the real and synthetically generated aircraft feature 10 fine grain attributes including: aircraft length, wingspan, wing-shape, wing-position, wingspan class, propulsion, number of engines, number of vertical-stabilizers, presence of canards, and aircraft role. Finally, we conduct extensive experiments to evaluate the real and synthetic datasets and compare performances. By doing so, we show the value of synthetic data for the task of detecting and classifying aircraft from an overhead perspective.

****

Spatially Aware Metadata for Raw Reconstruction
Abhijith Punnappurath, Michael S. Brown; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 218-226
A camera sensor captures a raw-RGB image that is then processed to a standard RGB (sRGB) image through a series of onboard operations performed by the camera's image signal processor (ISP). Among these processing steps, local tone mapping is one of the most important operations used to enhance the overall appearance of the final rendered sRGB image. For certain applications, it is often desirable to de-render or unprocess the sRGB image back to its original raw-RGB values. This "raw reconstruction" is a challenging task because many of the operations performed by the ISP, including local tone mapping, are nonlinear and difficult to invert. Existing raw reconstruction methods that store specialized metadata at capture time to enable raw recovery ignore local tone mapping and assume that a global transformation exists between the raw-RGB and sRGB color spaces. In this work, we advocate a spatially aware metadata-based raw reconstruction method that is robust to local tone mapping, and yields significantly higher raw reconstruction accuracy (6 dB average PSNR improvement) compared to existing raw reconstruction methods. Our method requires only 0.2% samples of the full-sized image as metadata, has negligible computational overhead at capture time, and can be easily integrated into modern ISPs.

****

3DPoseLite: A Compact 3D Pose Estimation Using Node Embeddings
Meghal Dani, Karan Narain, Ramya Hebbalaguppe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1878-1887
Efficient pose estimation finds utility in Augmented Reality (AR) and other computer vision applications such as autonomous navigation and robotics, to name a few. A compact and accurate pose estimation methodology is of paramount importance for on-device inference in such applications. Our proposed solution 3DPoseLite, estimates pose of generic objects by utilizing a compact node embedding representation, unlike computationally expensive multi-view and point-cloud representations. The neural network outputs a 3D pose, taking RGB image and its corresponding graph (obtained by skeletonizing the 3D meshes) as inputs. Our approach utilizes node2vec framework to learn low-dimensional representations for nodes in a graph by optimizing a neighborhood preserving objective. We achieve a space and time reduction by a factor of 11x and 3x respectively, with respect to the state-of-the-art approach, PoseFromShape, on benchmark Pascal3D dataset. We also test the performance of our model on unseen data using Pix3D dataset.

****

Synthetic Expressions Are Better Than Real for Learning to Detect Facial Actions
Koichiro Niinuma, Itir Onal Ertugrul, Jeffrey F. Cohn, Laszlo A. Jeni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1248-1257
Critical obstacles in training classifiers to detect facial actions are the limited sizes of annotated video databases and the relatively low frequencies of occurrence of many actions. To address these problems, we propose an approach that makes use of facial expression generation. Our approach reconstructs the 3D shape of the face from each video frame, aligns the 3D mesh to a canonical view, and then trains a GAN-based network to synthesize novel images with facial action units of interest. To evaluate this approach, a deep neural network was trained on two separate datasets: One network was trained on video of synthesized facial expressions generated from FERA17; the other network was trained on unaltered video from the same database. Both networks used the same train and validation par

titions and were tested on the test partition of actual video from FERA17. The network trained on synthesized facial expressions outperformed the one trained on actual facial expressions and surpassed current state-of-the-art approaches.

****

LoGAN: Latent Graph Co-Attention Network for Weakly-Supervised Video Moment Retrieval
Reuben Tan, Huijuan Xu, Kate Saenko, Bryan A. Plummer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2083-2092

The goal of weakly-supervised video moment retrieval is to localize the video segment most relevant to a description without access to temporal annotations during training. Prior work uses co-attention mechanisms to understand relationships between the vision and language data, but they lack contextual information between video frames that can be useful to determine how well a segment relates to the query. To address this, we propose an efficient Latent Graph Co-Attention Network (LoGAN) that exploits fine-grained frame-by-word interactions to jointly reason about the correspondences between all possible pairs of frames, providing context cues absent in prior work. Experiments on the DiDeMo and Charades-STA datasets demonstrate the effectiveness of our approach, where we improve Recall@1 by 5-20% over prior weakly-supervised methods, even boasting an 11% gain over strongly-supervised methods on DiDeMo, while also using significantly fewer model parameters than other co-attention mechanisms.

****

Representation Learning From Videos In-the-Wild: An Object-Centric Approach
Rob Romijnders, Aravindh Mahendran, Michael Tschannen, Josip Djolonga, Marvin Ritter, Neil Houlsby, Mario Lucic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 177-187

We propose a method to learn image representations from uncurated videos. We combine a supervised loss from off-the-shelf object detectors and self-supervised losses which naturally arise from the video-shot-frame-object hierarchy present in each video. We report competitive results on 19 transfer learning tasks of the Visual Task Adaptation Benchmark (VTAB), and on 8 out-of-distribution-generalization tasks, and discuss the benefits and shortcomings of the proposed approach. In particular, it improves over the baseline on all 18/19 few-shot learning tasks and 8/8 out-of-distribution generalization tasks. Finally, we perform several ablation studies and analyze the impact of the pretrained object detector on the performance across this suite of tasks.

****

Noisy Concurrent Training for Efficient Learning Under Label Noise
Fahad Sarfraz, Elahe Arani, Bahram Zonooz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3159-3168

Deep neural networks (DNNs) fail to learn effectively under label noise and have been shown to memorize random labels which affect their generalization performance. We consider learning in isolation, using one-hot encoded labels as the sole source of supervision, and a lack of regularization to discourage memorization as the major shortcomings of the standard training procedure. Thus, we propose Noisy Concurrent Training (NCT) which leverages collaborative learning to use the consensus between two models as an additional source of supervision. Furthermore, inspired by trial-to-trial variability in the brain, we propose a counterintuitive regularization technique, target variability, which entails randomly changing the labels of a percentage of training samples in each batch as a deterrent to memorization and overgeneralization in DNNs. Target variability is applied independently to each model to keep them diverged and avoid the confirmation bias. As DNNs tend to prioritize learning simple patterns first before memorizing the noisy labels, we employ a dynamic learning scheme whereby as the training progresses, the two models increasingly rely more on their consensus. NCT also progressively increases the target variability to avoid memorization in later stages. We demonstrate the effectiveness of our approach on both synthetic and real-world noisy benchmark datasets.

****

Dual-Stream Fusion Network for Spatiotemporal Video Super-Resolution

Min-Yuan Tseng, Yen-Chung Chen, Yi-Lun Lee, Wei-Sheng Lai, Yi-Hsuan Tsai, Wei-Chen Chiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2684-2693

Upsampling toward visual data has long been an important research topic for improving the perceptual quality and benefiting various computer vision applications. In recent years, we have witnessed remarkable progresses brought by the renaissance of deep learning techniques for video or image super-resolution. However, most existing works focus on advancing super-resolution at either spatial or temporal direction, i.e, to increase the spatial resolution or video frame rate. In this paper, we instead turn to discuss both directions jointly and tackle the spatiotemporal upsampling problem. Our method is based on an important observation that: even the direct cascade of prior researches in spatial and temporal super-resolution can achieve the spatiotemporal upsampling, different orders for combining them will lead to results with a complementary property. Thus, we propose a dual-stream fusion network to adaptively fuse the intermediate results produced by two spatiotemporal upsampling streams, where the first stream applies the spatial super-resolution followed by the temporal super-resolution, while the second one is with the reverse order of cascade. Extensive experiments verify the efficacy of the proposed model and its superior performance with respect to several baselines. Moreover, the investigation on utilizing various spatial and temporal upsampling methods as the basis in our two streams well demonstrates the flexibility and wide applicability of the proposed framework.

****

Automatic Quantification of Plant Disease From Field Image Data Using Deep Learning

Kanish Garg, Swati Bhugra, Brejesh Lall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1965-1972

Plant disease is a major factor in yield reduction. Thus, plant breeders currently rely on selecting disease-resistant plant cultivars, which involves disease severity rating of a large variety of cultivars. Traditional visual screening of these cultivars is an error-prone process, which necessitates the development of an automatic framework for disease quantification based on field-acquired images using unmanned aerial vehicles (UAVs) to augment the throughput. Since these images are impaired by complex backgrounds, uneven lighting, and densely overlapping leaves, state-of-the-art frameworks formulate the processing pipeline as a dichotomy problem (i.e. presence/absence of disease). However, additional information regarding accurate disease localization and quantification is crucial for breeders. This paper proposes a deep framework for simultaneous segmentation of individual leaf instances and corresponding diseased region using a unified feature map with a multi-task loss function for an end-to-end training. We test the framework on field maize dataset with Northern Leaf Blight (NLB) disease and the experimental results show a disease severity correlation of 73% with the manual ground truth data and run-time efficiency of 5fps.

****

Scale Equivariance Improves Siamese Tracking

Ivan Sosnovik, Artem Moskalev, Arnold W.M. Smeulders; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2765-2774

Siamese trackers turn tracking into similarity estimation between a template and the candidate regions in the frame. Mathematically, one of the key ingredients of success of the similarity function is translation equivariance. Non-translation-equivariant architectures induce a positional bias during training, so the location of the target will be hard to recover from the feature space. In real life scenarios, objects undergo various transformations other than translation, such as rotation or scaling. Unless the model has an internal mechanism to handle them, the similarity may degrade. In this paper, we focus on scaling and we aim to equip the Siamese network with additional built-in scale equivariance to capture the natural variations of the target a priori. We develop the theory for scale-equivariant Siamese trackers, and provide a simple recipe for how to make a wi de range of existing trackers scale-equivariant. We present SE-SiamFC, a scale-equivariant variant of SiamFC built according to the recipe. We conduct experiments on OTB and VOT benchmarks and on the synthetically generated T-MNIST and S-MNIST datasets. We demonstrate that a built-in additional scale equivariance is useful for visual object tracking.

****

On the Generalization of Learning-Based 3D Reconstruction

Miguel Angel Bautista, Walter Talbott, Shuangfei Zhai, Nitish Srivastava, Joshua M. Susskind; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2180-2189

State-of-the-art learning-based monocular 3D reconstruction methods learn priors over object categories on the training set, and as a result struggle to achieve reasonable generalization to object categories unseen during training. In this paper we study the inductive biases encoded in the model architecture that impact the generalization of learning-based 3D reconstruction methods. We find that 3 inductive biases impact performance: the spatial extent of the encoder, the use of the underlying geometry of the scene to describe point features, and the mechanism to aggregate information from multiple views. Additionally, we propose mechanisms to enforce those inductive biases: a point representation that is aware of camera position, and a variance cost to aggregate information across views. Our model achieves state-of-the-art results on the standard ShapeNet 3D reconstruction benchmark in various settings.

****

Learning Data Augmentation With Online Bilevel Optimization for Image Classification

Saypraseuth Mounsaveng, Issam Laradji, Ismail Ben Ayed, David Vazquez, Marco Pedersoli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1691-1700

Data augmentation is a key practice in machine learning for improving generalization performance. However, finding the best data augmentation hyperparameters requires domain knowledge or a computationally demanding search. We address this issue by proposing an efficient approach to automatically train a network that learns an effective distribution of transformations to improve its generalization. Using bilevel optimization, we directly optimize the data augmentation parameters using a validation set. This framework can be used as a general solution to learn the optimal data augmentation jointly with an end task model like a classifier. Results show that our joint training method produces an image classification accuracy that is comparable to or better than carefully hand-crafted data augmentation. Yet, it does not need an expensive external validation loop on the data augmentation hyperparameters.

****

Seeing Through Your Skin: Recognizing Objects With a Novel Visuotactile Sensor

Francois R. Hogan, Michael Jenkin, Sahand Rezaei-Shoshtari, Yogesh Girdhar, David Meger, Gregory Dudek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1218-1227

We introduce a new class of vision-based sensor and associated algorithmic processes that combine visual imaging with high-resolution tactile sending, all in a uniform hardware and computational architecture. We demonstrate the sensor's efficacy for both multi-modal object recognition and metrology. Object recognition is typically formulated as an unimodal task, but by combining two sensor modalities we show that we can achieve several significant performance improvements. This sensor, named the See-Through-your-Skin sensor (STS), is designed to provide rich multi-modal sensing of contact surfaces. Inspired by recent developments in optical tactile sensing technology, we address a key missing feature of these sensors: the ability to capture a visual perspective of the region beyond the contact surface. Whereas optical tactile sensors are typically opaque, we present a sensor with a semitransparent skin that has the dual capabilities of acting as a tactile sensor and/or as a visual camera depending on its internal lighting conditions. This paper details the design of the sensor, showcases its dual sensing capabilities, and presents a deep learning architecture that fuses vision and touch. We validate the ability of the sensor to classify household objects, recognize fine textures, and infer their physical properties both through numerical simulations and experimentally with a smart countertop prototype.

****

Audio-Visual Event Localization via Recursive Fusion by Joint Co-Attention

Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, Yan Yan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4013-4022

The major challenge in audio-visual event localization task lies in how to fuse information from multiple modalities effectively. Recent works have shown that the attention mechanism is beneficial to the fusion process. In this paper, we propose a novel joint attention mechanism with multimodal fusion methods for audio-visual event localization. Particularly, we present a concise yet valid architecture that effectively learns representations from multiple modalities in a joint manner. Initially, visual features are combined with auditory features and then turned into joint representations. Next, we make use of the joint representations to attend to visual features and auditory features, respectively. With the help of this joint co-attention, new visual and auditory features are produced, and thus both features can enjoy the mutually improved benefits from each other. It is worth noting that the joint co-attention unit is recursive meaning that it can be performed multiple times for obtaining better joint representations progressively. Extensive experiments on the public AVE dataset have shown that the proposed method achieves significantly better results than the state-of-the-art methods.

****

SWAG: Superpixels Weighted by Average Gradients for Explanations of CNNs

Thomas Hartley, Kirill Sidorov, Christopher Willis, David Marshall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 423-432

Providing an explanation of the operation of CNNs that is both accurate and interpretable is becoming essential in fields like medical image analysis, surveillance, and autonomous driving. In these areas, it is important to have confidence that the CNN is working as expected and explanations from saliency maps provide an efficient way of doing this. In this paper, we propose a pair of complementary contributions that improve upon the state of the art for region-based explanations in both accuracy and utility. The first is SWAG, a method for generating accurate explanations quickly using superpixels for discriminative regions which is meant to be a more accurate, efficient, and tunable drop in replacement method for Grad-CAM, LIME, or other region-based methods. The second contribution is based on an investigation into how to best generate the superpixels used to represent the features found within the image. Using SWAG, we compare using superpixels created from the image, a combination of the image and backpropagated gradients, and the gradients themselves. To the best of our knowledge, this is the first method proposed to generate explanations using superpixels explicitly created to represent the discriminative features important to the network. To compare we use both ImageNet and challenging fine-grained datasets over a range of metrics . We demonstrate experimentally that our methods provide the best local and global accuracy compared to Grad-CAM, Grad-CAM++, LIME, XRAI, and RISE.

****

Keypoint-Aligned Embeddings for Image Retrieval and Re-Identification

Olga Moskvyak, Frederic Maire, Feras Dayoub, Mahsa Baktashmotlagh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021 , pp. 676-685

Learning embeddings that are invariant to the pose of the object is crucial in visual image retrieval and re-identification. The existing approaches for person, vehicle, or animal re-identification tasks suffer from high intra-class variance due to deformable shapes and different camera viewpoints. To overcome this limitation, we propose to align the image embedding with a predefined order of the keypoints. The proposed keypoint aligned embeddings model (KAE-Net) learns part-level features via multi-task learning which is guided by keypoint locations. Mo re specifically, KAE-Net extracts channels from a feature map activated by a specific keypoint through learning the auxiliary task of heatmap reconstruction for this keypoint. The KAE-Net is compact, generic and conceptually simple. It achieves state of the art performance on the benchmark datasets of CUB-200-2011, Cars196 and VeRi-776 for retrieval and re-identification tasks.

****

HealTech - A System for Predicting Patient Hospitalization Risk and Wound Progression in Old Patients

Subba Reddy Oota, Vijay Rowtula, Shahid Mohammed, Jeffrey Galitz, Minghsun Liu, Manish Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2463-2472

How bad is my wound? How fast will the wound heal? Do I need to get hospitalized? Questions like these are critical for wound assessment, but challenging to answer. Given a wound image and patient attributes, our goal is to build models for two wound assessment tasks: (1) predicting if the patient needs hospitalization for the wound to heal, and (2) estimating wound progression, i.e., weeks to heal. The problem is challenging because wound progression and hospitalization risk depend on multiple factors that need to be inferred automatically from the given wound image. There exists no work which performs a rigorous study of wound assessment tasks considering multiple wound attributes inferred using a large dataset of wound images. We present HealTech, a two-stage wound assessment solution. The first stage predicts various wound attributes (like ulcer type, location, stage, etc.) from wound images, using deep neural networks. The second stage predicts (1) whether the wound would heal (using conventional in-house treatment) or not (needs hospitalization), and (2) the number of weeks to heal, using an evolutionary algorithm based stacked Light Gradient Boosted Machines (LGBM) model. On a large dataset of 125711 wound images, HealTech achieves a recall of 83 and a precision of 92 for wounds with the risk of hospitalization. For wounds that can be healed without hospitalization, precision and recall are as high as 99. Our wound progression model provides a mean absolute error of 3.3 weeks.

****

A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset

Domenick Poster, Matthew Thielke, Robert Nguyen, Srinivasan Rajaraman, Xing Di, Cedric Nimpa Fondje, Vishal M. Patel, Nathaniel J. Short, Benjamin S. Riggan, Nasser M. Nasrabadi, Shuowen Hu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1559-1568

Thermal face imagery, which captures the naturally emitted heat from the face, is limited in availability compared to face imagery in the visible spectrum. To help address this scarcity of thermal face imagery for research and algorithm development, we present the DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF). With over 500,000 images from 395 subjects, the ARL-VTF dataset represents, to the best of our knowledge, the largest collection of paired visible and thermal face images to date. The data was captured using a modern long wave infrared (LWIR) camera mounted alongside a stereo setup of three visible spectrum cameras. Variability in expressions, pose, and eyewear has been systematically recorded. The dataset has been curated with extensive annotations, metadata, and standardized protocols for evaluation. Furthermore, this paper presents extensive benchmark results and analysis on thermal face landmark detection and thermal-to-visible face verification by evaluating state-of-the-art models on the ARL-VTF dataset.

****

RGPNet: A Real-Time General Purpose Semantic Segmentation

Elahe Arani, Shabbir Marzban, Andrei Pata, Bahram Zonooz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3009-3018

We propose a real-time general purpose semantic segmentation architecture, RGPNet, which achieves significant performance gain in complex environments. RGPNet consists of a light-weight asymmetric encoder-decoder and an adaptor. The adaptor helps preserve and refine the abstract concepts from multiple levels of distributed representations between encoder and decoder. It also facilitates the gradie nt flow from deeper layers to shallower layers. Our experiments demonstrate that RGPNet can generate segmentation results in real-time with comparable accuracy to the state-of-the-art non-real-time heavy models. Moreover, towards green AI, we show that using an optimized label-relaxation technique with progressive resizing can reduce the training time by up to 60% while preserving the performance. We conclude that RGPNet obtains a better speed-accuracy trade-off across multiple datasets.

****

Style Transfer by Rigid Alignment in Neural Net Feature Space

Suryabhan Singh Hada, Miguel A. Carreira-Perpinan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2576-2585

Arbitrary style transfer is an important problem in computer vision that aims to transfer style patterns from an arbitrary style image to a given content image. However, current methods either rely on slow iterative optimization or fast pre-determined feature transformation, but at the cost of compromised visual quality of the styled image; especially, distorted content structure. In this work, we present an effective and efficient approach for arbitrary style transfer that seamlessly transfers style patterns as well as keep content structure intact in the styled image. We achieve this by aligning style features to content features using rigid alignment; thus modifying style features, unlike the existing methods that do the opposite. We demonstrate the effectiveness of the proposed approach by generating high-quality stylized images and compare the results with the current state-of-the-art techniques for arbitrary style transfer.

****

Confidence-Driven Hierarchical Classification of Cultivated Plant Stresses

Logan Frank, Christopher Wiegman, Jim Davis, Scott Shearer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2503-2512

The application of convolutional neural networks (CNNs) and deep learning to different domains has become increasingly popular in the last several years. In particular, such models have been used in the agriculture domain to identify plant species, identify plant stresses, and estimate crop yields. Although there has been much success in applying these techniques to the agriculture domain, these works contain many shortcomings that are hindering their chance for adoption in practice (e.g., lack of domain knowledge, predicting only specific stress types, etc.). We address issues of previous works for the task of plant stress identification by applying a hierarchical classification approach employing confidence as a means to determine the specificity of a classification. This work is a collaboration between computer science and agricultural engineering experts.

****

ADA-AT/DT: An Adversarial Approach for Cross-Domain and Cross-Task Knowledge Transfer

Ruchika Chavhan, Ankit Jha, Biplab Banerjee, Subhasis Chaudhuri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3502-3511

We deal with the problem of cross-task and cross-domain knowledge transfer in the realm of scene understanding for autonomous vehicles. We consider the scenario where supervision is available for a pair of tasks in a source domain while it is available for only one of the tasks in the target domain. Given that, the goal is to perform inference for the task in the target which is devoid of any training information. We argue that the only reported work in learning across tasks and domains (AT/DT) [23] faces the problem of domain shift between the source and target domains, hindering predictions on the target domain when the transfer of knowledge is learned on a statistically different yet related source domain. As a remedy, we develop a novel framework called ADA-AT/DT based on the adversarial training strategy to ensure that the domain-gaps are minimized for the common cross-domain supervised task. This, in effect, helps in realizing a domain-independent task-transfer function that eventually helps in performing improved inference in the target domain. We demonstrate that our proposed method significantly outperforms [23] by using models with 81% fewer trainable parameters. In addit ion, we perform experiments on a transformation mapping similar to U-Net to ensure maximum exploitation of features for task transfer. Extensive experiments have been performed on four different domains (Synthia, CityScapes, Carla, and KITTI) for two visual tasks (depth estimation and semantic segmentation) to confirm the superiority of our method.

****

Data-Free Knowledge Distillation for Object Detection

Akshay Chawla, Hongxu Yin, Pavlo Molchanov, Jose Alvarez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3289-3298

We present DeepInversion for Object Detection (DIODE) to enable data-free knowledge distillation for neural networks trained on the object detection task. From a data-free perspective, DIODE synthesizes images given only an off-the-shelf pre-trained detection network and without any prior domain knowledge, generator network, or pre-computed activations. DIODE relies on two key components--first, an extensive set of differentiable augmentations to improve image fidelity and distillation effectiveness. Second, a novel automated bounding box and category sampling scheme for image synthesis enabling generating a large number of images with a diverse set of spatial and category objects. The resulting images enable data-free knowledge distillation from a teacher to a student detector, initialized from scratch. In an extensive set of experiments, we demonstrate that DIODE's ability to match the original training distribution consistently enables more effective knowledge distillation than out-of-distribution proxy datasets, which unavoidably occur in a data-free setup given the absence of the original domain knowledge.

****

Class Anchor Clustering: A Loss for Distance-Based Open Set Recognition

Dimity Miller, Niko Sunderhauf, Michael Milford, Feras Dayoub; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3570-3578

In open set recognition, deep neural networks encounter object classes that were unknown during training. Existing open set classifiers distinguish between known and unknown classes by measuring distance in a network's logit space, assuming that known classes cluster closer to the training data than unknown classes. However, this approach is applied post-hoc to networks trained with cross-entropy loss, which does not guarantee this clustering behaviour. To overcome this limitation, we introduce the Class Anchor Clustering (CAC) loss. CAC is a distance-based loss that explicitly trains known classes to form tight clusters around anchored class-dependent centres in the logit space. We show that training with CAC achieves state-of-the-art performance for distance-based open set classifiers on all six standard benchmark datasets, with a 15.2% AUROC increase on the challenging TinyImageNet, without sacrificing classification accuracy. We also show that our anchored class centres achieve higher open set performance than learnt class centres, particularly on object-based datasets and large numbers of training classes.

****

Deep Active Learning for Joint Classification & Segmentation With Weak Annotator

Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3338-3347

CNN visualization and interpretation methods, like class-activation maps (CAMs), are typically used to highlight the image regions linked to class predictions. These models allow to simultaneously classify images and extract class-dependent saliency maps, without the need for costly pixel-level annotations. However, they typically yield segmentations with high false-positive rates and, therefore, coarse visualisations, more so when processing challenging images, as encountered in histology. To mitigate this issue, we propose an active learning (AL) framework, which progressively integrates pixel-level annotations during training. Given training data with global image-level labels, our deep weakly-supervised learning model jointly performs supervised image-level classification and active le arning for segmentation, integrating pixel annotations by an oracle. Unlike standard AL methods that focus on sample selection, we also leverage large numbers of unlabeled images via pseudo-segmentations (i.e., self-learning at the pixel level), and integrate them with the oracle-annotated samples during training. We report extensive experiments over two challenging benchmarks -- high-resolution medical images (histology GlaS data for colon cancer) and natural images (CUB-200-2011 for bird species). Our results indicate that, by simply using random sample selection, the proposed approach can significantly outperform state-of the-art CAMs and AL methods, with an identical oracle-supervision budget. Our code is publicly available.

****

Dynamic Plane Convolutional Occupancy Networks

Stefan Lionar, Daniil Emtsev, Dusan Svilarkovic, Songyou Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1829-1838

Learning-based 3D reconstruction using implicit neural representations has shown promising progress not only at the object level but also in more complicated scenes. In this paper, we propose Dynamic Plane Convolutional Occupancy Networks, a novel implicit representation pushing further the quality of 3D surface reconstruction. The input noisy point clouds are encoded into per-point features that are projected onto multiple 2D dynamic planes. A fully-connected network learns to predict plane parameters that best describe the shapes of objects or scenes. To further exploit translational equivariance, convolutional neural networks are applied to process the plane features. Our method shows superior performance in surface reconstruction from unoriented point clouds in ShapeNet as well as an indoor scene dataset. Moreover, we also provide interesting observations on the distribution of learned dynamic planes.

****

Fine-Grained Foreground Retrieval via Teacher-Student Learning

Zongze Wu, Dani Lischinski, Eli Shechtman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3646-3654

Foreground image retrieval is a fundamental task in computer vision. Given an image of the background scene with a bounding box indicating the target location, the goal is to retrieve a set of images of foreground objects from a given category, which are semantically compatible with the background. We formulate foreground retrieval as a self-supervised domain adaptation task, where the source domain consists of foreground images and the target domain of background images. Specifically, given pretrained object feature extraction networks that serve as teachers, we train a student network to infer compatible foreground features from background images. Thus, foregrounds and backgrounds are effectively mapped into a common feature space, enabling retrieval of the foregrounds that are closest to the target background in that space. A notable feature of our approach is that our training strategy does not require instance segmentation, unlike current state-of-the-art methods. Thus, our method may be applied to diverse foreground categories and background scene types and enables us to retrieve the foreground in a fine-grained manner, which is closer to the requirements of real world applications.

****

Unsupervised Multimodal Video-to-Video Translation via Self-Supervised Learning

Kangning Liu, Shuhang Gu, Andres Romero, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1030-1040

Existing unsupervised video-to-video translation methods fail to produce translated videos which are frame-wise realistic, semantic information preserving and video-level consistent. In this work, we propose a novel unsupervised video-to-video translation model. Our model decomposes the style and the content uses the specialized encoder-decoder structure and propagates the inter-frame information through bidirectional recurrent neural network (RNN) units. The style-content decomposition mechanism enables us to achieve style-consistent video translation results as well as provides us with a good interface for modality flexible transl ation. In addition, by changing the input frames and style codes incorporated in our translation, we propose a video interpolation loss, which captures temporal information within the sequence to train our building blocks in a self-supervised manner. Our model can produce photo-realistic, spatio-temporal consistent translated videos in a multimodal way. Subjective and objective experimental results validate the superiority of our model over existing methods.

****

Kernel Self-Attention for Weakly-Supervised Image Classification Using Deep Multiple Instance Learning

Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, Bartosz Zielinski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1721-1730

Not all supervised learning problems are described by a pair of a fixed-size input tensor and a label. In some cases, especially in medical image analysis, a label corresponds to a bag of instances (e.g. image patches), and to classify such bag, aggregation of information from all of the instances is needed. There have been several attempts to create a model working with a bag of instances, however, they are assuming that there are no dependencies within the bag and the label is connected to at least one instance. In this work, we introduce Self-Attention Attention-based MIL Pooling (SA-AbMILP) aggregation operation to account for the dependencies between instances. We conduct several experiments on MNIST, histological, microbiological, and retinal databases to show that SA-AbMILP performs better than other models. Additionally, we investigate kernel variations of Self-Attention and their influence on the results.

****

Real-Time Gait-Based Age Estimation and Gender Classification From a Single Image

Chi Xu, Yasushi Makihara, Ruochen Liao, Hirotaka Niitsuma, Xiang Li, Yasushi Yagi, Jianfeng Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3460-3470

In this paper, we propose a unified real-time framework for gait-based age estimation and gender classification that uses just a single image, which reduces the latency in video capturing compared with the existing methods based on a gait cycle. To cope with the problem of lacking motion information in the input single image, we first reconstruct a gait cycle of a silhouette sequence from the input image via a gait cycle reconstruction network. The reconstructed gait cycle is then fed into a state-of-the-art gait recognition network for feature representation learning, which is further used to obtain the class of the gender and the estimated probability distribution of integer age labels. Unlike the existing methods focusing on the gait sequences captured from the side view, the proposed method is applicable to the gait images from an arbitrary view with a single trained model, which is more suitable for real-world application scenarios (e.g., automatic access control). Stand-alone and client-server online systems were implemented based on the proposed method, which validates the real-time/online property in actual scenes. The experiments on the world's largest multi-view gait dataset demonstrate the effectiveness of the proposed method, which achieves performance improvement compared with the benchmark algorithms.

****

Generative Patch Priors for Practical Compressive Image Recovery

Rushil Anirudh, Suhas Lohit, Pavan Turaga; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2535-2545

In this paper, we propose the generative patch prior (GPP) that defines a generative prior for compressive image recovery, based on patch-manifold models. Unlike learned, image-level priors that are restricted to the range space of a pre-trained generator, GPP can recover a wide variety of natural images using a pre-trained patch generator. Additionally, GPP retains the benefits of generative priors like high reconstruction quality at extremely low sensing rates, while also being much more generally applicable. We show that GPP outperforms several unsupervised and supervised techniques on three different sensing models -- linear compressive sensing with known, and unknown calibration settings, and the non-linea r phase retrieval problem. Finally, we propose an alternating optimization strategy using GPP for joint calibration-and-reconstruction which performs favorably against several baselines on a real world, un-calibrated compressive sensing dataset.

****

ExMaps: Long-Term Localization in Dynamic Scenes Using Exponential Decay

Alexandros Rotsidis, Christof Lutteroth, Peter Hall, Christian Richardt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2867-2876

Visual camera localization using offline maps is widespread in robotics and mobile applications. Most state-of-the-art localization approaches assume static scenes, so maps are often reconstructed once and then kept constant. However, many scenes are dynamic and as changes in the scene happen, future localization attempts may struggle or fail entirely. Therefore, it is important for successful long-term localization to update and maintain maps as new observations of the scene, and changes in it, arrive. We propose a novel method for automatically discovering which points in a map remain stable over time, and which are due to transient changes. To this end, we calculate a stability store for each point based on its visibility over time, weighted by an exponential decay over time. This allows us to introduce the impact of time when scoring points, and distinguishes which points are useful for long-term localization. We evaluate our method on the CMU Extended Seasons dataset (outdoors) and a new indoor dataset of a retail shop, and show the benefit of maintaining a `live map' that integrates updates over time using our exponential decay based method over a static `base map'.

****

FACEGAN: Facial Attribute Controllable rEenactment GAN

Soumya Tripathy, Juho Kannala, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1329-1338

The face reenactment is a popular facial animation method where the person's identity is taken from the source image and the facial motion from the driving image. Recent works have demonstrated high-quality results by combin- ing the facial landmark-based motion representations with the generative adversarial networks. These models perform best if the source and driving images depict the same person or if the facial structures are otherwise very similar. However, if the identity differs, the driving facial structures leak to the output distorting the reenactment result. We propose a novel Facial Attribute Controllable rEenactment GAN (FACEGAN), which transfers the facial motion from the driving face via the Action Unit (AU) representation. Unlike facial landmarks, the AUs are independent of the facial structure preventing the identity leak. Moreover, AUs provide a human interpretable way to control the reenactment. FACEGAN processes background and face regions separately for optimized output quality. The extensive quantitative and qualitative comparisons show a clear improvement over the state-of-the-art in a single source reenactment task. The results are best illustrated in the reenactment video provided in the supplementary material. The source code will be made available upon publication of the paper.

****

Effective Fusion Factor in FPN for Tiny Object Detection

Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, Zhenjun Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1160-1168

FPN-based detectors have made significant progress in general object detection,e.g., MS COCO and CityPersons.However, these detectors fail in certain application scenarios,e.g., tiny object detection. In this paper, we argue that the top-down connections between adjacent layers in FPN bring two-side influences for tiny object detection, not only positive. We propose a novel concept, fusion factor, to control information that deep layers deliver to shallow layers,for adapting FPN to tiny object detection. After series of experiments and analysis, we explore how to estimate an effective value of fusion factor for a particular dataset by a statistical method. The estimation is dependent on the number of objects distributed to each layer. Comprehensive experiments are conducted on tiny object

detection datasets,e.g., TinyPerson and Tiny CityPersons. Our results show that when configuring FPN with a proper fusion factor, the network is able to achieve significant performance gains over the baseline on tiny object detection datasets. Codes and models will be released.
****
Deep Interactive Thin Object Selection
Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, Jiashi Feng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 305-314
Existing deep learning based interactive segmentation methods have achieved remarkable performance with only a few user clicks, e.g. DEXTR attaining 91.5% IoU on PASCAL VOC with only four extreme clicks. However, we observe even the state-of-the-art methods would often struggle in cases of objects to be segmented with elongated thin structures (e.g. bug legs and bicycle spokes). We investigate such failures, and find the critical reasons behind are two-fold: 1) lack of appropriate training dataset; and 2) extremely imbalanced distribution w.r.t. number of pixels belonging to thin and non-thin regions. Targeted at these challenges, we collect a large-scale dataset specifically for segmentation of thin elongated objects, named ThinObject-5K. Also, we present a novel integrative thin object segmentation network consisting of three streams. Among them, the high-resolution edge stream aims at preserving fine-grained details including elongated thin parts; the fixed-resolution context stream focuses on capturing semantic contexts. The two streams' outputs are then amalgamated in the fusion stream to complement each other for help producing a refined segmentation output with sharper predictions around thin parts. Extensive experimental results well demonstrate the effectiveness of our proposed solution on segmenting thin objects, surpassing the baseline by 30% IoU_thin despite using only four clicks. Codes and dataset are available at https://github.com/liewjunhao/thin-object-selection.
****
Zero-Pair Image to Image Translation Using Domain Conditional Normalization
Samarth Shukla, Andres Romero, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3512-3519
In this paper, we propose an approach based on domain conditional normalization (DCN) for zero-pair image-to-image translation, i.e., translating between two domains which have no paired training data available but each have paired training data with a third domain. We employ a single generator which has an encoder-decoder structure and analyze different implementations of domain conditional normalization to obtain the desired target domain output. The validation benchmark uses RGB-depth pairs and RGB-semantic pairs for training and compares performance for the depth-semantic translation task. The proposed approaches improve in qualitative and quantitative terms over the compared methods, while using much fewer parameters.
****
RNNP: A Robust Few-Shot Learning Approach
Pratik Mazumder, Pravendra Singh, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2664-2673
Learning from a few examples is an important practical aspect of training classifiers. Various works have examined this aspect quite well. However, all existing approaches assume that the few examples provided are always correctly labeled. This is a strong assumption, especially if one considers the current techniques for labeling using crowd-based labeling services. We address this issue by proposing a novel robust few-shot learning approach. Our method relies on generating robust prototypes from a set of few examples. Specifically, our method refines the class prototypes by producing hybrid features from the support examples of each class. The refined prototypes help to classify the query images better. Our method can replace the evaluation phase of any few-shot learning method that uses a nearest neighbor prototype-based evaluation procedure to make them robust. We evaluate our method on standard mini-ImageNet and tiered-ImageNet datasets. We

perform experiments with various label corruption rates in the support examples of the few-shot classes. We obtain significant improvement over widely used few-shot learning methods that suffer significant performance degeneration in the presence of label noise. We finally provide extensive ablation experiments to validate our method.

****

An Alternative of LIDAR in Nighttime: Unsupervised Depth Estimation Based on Single Thermal Image

Yawen Lu, Guoyu Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3833-3843

Most existing autonomous driving vehicles and robots rely on active LIDAR sensors to detect the depth of the surrounding environment, which usually has limited resolution, and the emitted laser can be harmful to people and the environment. Current passive image-based depth estimation algorithms focus on color images from RGB sensors, which is not suitable for dark and night environment with limited lighting resource. In this paper, we propose a framework to estimate the scene depth directly from a single thermal image that can still observe the scene in the low lighting condition. We learn the thermal image depth estimation framework together with RGB cameras, which also mitigates the training condition due to the easy availability of RGB cameras. With the translated thermal images from color images from our generative adversarial network, our depth estimation method can explore the unique characteristics in thermal images through our novel contour and edge-aware constraints to obtain a stable and anti-artifact disparity. We apply the commonly available color cameras to navigate the learning process of thermal image depth estimation framework. With our approach, an accurate depth map can be predicted without any prior knowledge under various illumination conditions. Experiments in public dataset, as well as our newly collected data, demonstrate superior performance of our method on single thermal image depth estimation compared with other state-of-the-art algorithms.

****

Class-Agnostic Few-Shot Object Counting

Shuo-Diao Yang, Hung-Ting Su, Winston H. Hsu, Wen-Chin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 870-878

Object counting which aims to calculate the number of total instances of the given class is a classic but crucial task that can be applied to many applications. Most of the prior works only focus on counting certain classes of objects such as people, cars, animals, etc. However, in recent years, there are lots of applications that need to get the count of the unseen class of objects such as a mechanical arm commanded to grab the novel object. In this paper, we present an effective object counting network, Class-agnostic Few-shot Object Counting Network (CFOCNet), that supports counting arbitrary classes of object unseen during training stage. Instead of counting a pre-defined class, our model is able to count instances based on input reference images and reduces the huge cost of data collection, training and parameter tuning for each new object class. Our model utilizes not only the similarity between query image and reference images but self attending the query image to learn the self-repeatedness. Using a two-stream Resnet that matches features in different scales, our network can automatically learn to aggregate different scales of the matching scores. We evaluate our method on the subset of the COCO dataset that contains 80 classes of objects and many diverse scenes. In the experiments, our network outperforms other methods including detection and some previous works by a large margin. To the best of our knowledge, we are the first that mainly focuses on few-shot object counting in the class-agnostic manner.

****

Self-Supervised Training for Blind Multi-Frame Video Denoising

Valery Dewil, Jeremy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, Pablo Arias; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2724-2734

We propose a self-supervised approach for training multi-frame video denoising n etworks. These networks predict frame t from a window of frames around t. Our self-supervised approach benefits from the video temporal consistency by penalizing a loss between the predicted frame t and a neighboring target frame, which are aligned using an optical flow. We use the proposed strategy for online internal learning, where a pre-trained network is fine-tuned to denoise a new unknown noise type from a single video. After a few frames, the proposed fine-tuning reaches and sometimes surpasses the performance of a state-of-the-art network trained with supervision. In addition, for a wide range of noise types, it can be applied blindly without knowing the noise distribution. We demonstrate this by showing results on blind denoising of different synthetic and real noises.

****

Improving Few-Shot Learning Using Composite Rotation Based Auxiliary Task

Pratik Mazumder, Pravendra Singh, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2654-2663

In this paper, we propose an approach to improve few-shot classification performance using a composite rotation based auxiliary task. Few-shot classification methods aim to produce neural networks that perform well for classes with a large number of training samples and classes with less number of training samples. They employ techniques to enable the network to produce highly discriminative features that are also very generic. Generally, the better the quality and generic-nature of the features produced by the network, the better is the performance of the network on few-shot learning. Our approach aims to train networks to produce such features by using a self-supervised auxiliary task. Our proposed composite rotation based auxiliary task performs rotation at two levels, i.e., rotation of patches inside the image (inner rotation) and rotation of the whole image (outer rotation) and assigns one out of 16 rotation classes to the modified image. We then simultaneously train for the composite rotation prediction task along with the original classification task, which forces the network to learn high-quality generic features that help improve the few-shot classification performance. We experimentally show that our approach performs better than existing few-shot learning methods on multiple benchmark datasets.

****

Covariance-Free Partial Least Squares: An Incremental Dimensionality Reduction Method

Artur Jordao, Maiko Lie, Victor Hugo Cunha de Melo, William Robson Schwartz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1421-1429

Dimensionality reduction plays an important role in computer vision problems since it reduces computational cost and is often capable of yielding more discriminative data representation. In this context, Partial Least Squares (PLS) has presented notable results in tasks such as image classification and neural network optimization. However, PLS is infeasible on large datasets (e.g., ImageNet) because it requires all the data to be in memory in advance, which is often impractical due to hardware limitations. Additionally, this requirement prevents us from employing PLS on streaming applications where the data are being continuously generated. Motivated by this, we propose a novel incremental PLS, named Covariance-free Incremental Partial Least Squares (CIPLS), which learns a low-dimensional representation of the data using a single sample at a time. In contrast to other state-of-the-art approaches, instead of adopting a partially-discriminative or SGD-based model, we extend Nonlinear Iterative Partial Least Squares (NIPALS) -- the standard algorithm used to compute PLS -- for incremental processing. Among the advantages of this approach are the preservation of discriminative information across all components, the possibility of employing its score matrices for feature selection, and its computational efficiency. We validate CIPLS on face verification and image classification tasks, where it outperforms several other incremental dimensionality reduction methods. In the context of feature selection, CIPLS achieves comparable results when compared to state-of-the-art techniques.

****

OverNet: Lightweight Multi-Scale Super-Resolution With Overscaling Network

Parichehr Behjati, Pau Rodriguez, Armin Mehri, Isabelle Hupont, Carles Fernandez Tena, Jordi Gonzalez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2694-2703

Super-resolution (SR) has achieved great success due to the development of deep convolutional neural networks (CNNs). However, as the depth and width of the networks increase, CNN-based SR methods have been faced with the challenge of computational complexity in practice. Moreover, most SR methods train a dedicated model for each target resolution, losing generality and increasing memory requirements. To address these limitations we introduce OverNet, a deep but lightweight convolutional network to solve SISR at arbitrary scale factors with a single model. We make the following contributions: first, we introduce a lightweight feature extractor that enforces efficient reuse of information through a novel recursive structure of skip and dense connections. Second, to maximize the performance of the feature extractor, we propose a model agnostic reconstruction module that generates accurate high-resolution images from overscaled feature maps obtained from any SR architecture. Third, we introduce a multi-scale loss function to achieve generalization across scales. Experiments show that our proposal outperforms previous state-of-the-art approaches in standard benchmarks, while maintaining relatively low computation and memory requirements.

****

One-Shot Image Recognition Using Prototypical Encoders With Reduced Hubness

Chenxi Xiao, Naveen Madapana, Juan Wachs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2252-2261

Humans have the innate ability to recognize new objects just by looking at sketches of them (also referred as to prototype images). Similarly, prototypical images can be used as an effective visual representations of unseen classes to tackle few-shot learning (FSL) tasks. Our main goal is to recognize unseen hand signs (gestures) traffic-signs, and corporate-logos, by having their iconographic images or prototypes. Previous works proposed to utilize variational prototypical-encoders (VPE) to address FSL problems. While VPE learns an image-to-image translation task efficiently, we discovered that its performance is significantly hampered by the so-called hubness problem and it fails to regulate the representations in the latent space. Hence, we propose a new model (VPE++) that inherently reduces hubness and incorporates contrastive and multi-task losses to increase the discriminative ability of FSL models. Results show that the VPE++ approach can generalize better to the unseen classes and can achieve superior accuracies on logos, traffic signs, and hand gestures datasets as compared to the state-of-the-art.

****

Real-Time RGBD-Based Extended Body Pose Estimation

Renat Bashirov, Anastasia Ianina, Karim Iskakov, Yevgeniy Kononenko, Valeriya Strizhkova, Victor Lempitsky, Alexander Vakhitov; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2807-2816

We present a system for real-time RGBD-based estimation of 3D human pose. We use parametric 3D deformable human mesh model (SMPL-X) as a representation and focus on the real-time estimation of parameters for the body pose, hands pose and facial expression from Kinect Azure RGB-D camera. We train estimators of body pose and facial expression parameters. Both estimators use previously published landmark extractors as input and custom annotated datasets for supervision, while hand pose is estimated directly by a previously published method. We combine the predictions of those estimators into a temporally-smooth human pose. We train the facial expression extractor on a large talking face dataset, which we annotate with facial expression parameters. For the body pose we collect and annotate a dataset of 56 people captured from a rig of 5 Kinect Azure RGB-D cameras and use it together with a large motion capture AMASS dataset. Our RGB-D body pose model outperforms the state-of-the-art RGB-only methods and works on the same level of accuracy compared to a slower RGB-D optimization-based solution. The combined system runs at 25 FPS on a server with a single GPU. The code will be available at saic-violet.github.io/rgbd-kinect-pose

****

Group Softmax Loss With Discriminative Feature Grouping

Takumi Kobayashi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2615-2624

In the supervised learning framework, a softmax cross-entropy loss is commonly applied to train deep neural networks for high-performance classification. It, however, demands large amount of annotated data and fails to learn the discriminative networks on a smaller amount of data. In this paper, we propose a novel loss measure to train the networks such that discriminative feature representation can be learned even on the smaller-scale dataset. By means of feature grouping, we effectively expose non-discriminative feature components to representation learning and formulate two types of group softmax losses to cope with the grouped features. The proposed method encourages discriminative representation across all feature components, and from a theoretical viewpoint it renders adversarial training which works for alleviating over-fitting especially on scarce training data. The experimental results on image classification tasks demonstrate that the proposed loss favorably improves performance of CNNs on various-scale data.

****

Separable Four Points Fundamental Matrix

Gil Ben-Artzi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 188-196

We present a novel approach for RANSAC-based computation of the fundamental matrix based on epipolar homography decomposition. We analyze the geometrical meaning of the decomposition-based representation and show that it directly induces a consecutive sampling strategy of two independent sets of correspondences. We show that our method guarantees a minimal number of evaluated hypotheses with respect to current minimal approaches, on the condition that there are four correspondences on an image line. We validate our approach on real-world image pairs, providing fast and accurate results.

****

Multimodal Trajectory Predictions for Autonomous Driving Without a Detailed Prior Map

Atsushi Kawasaki, Akihito Seki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3723-3732

Predicting the future trajectories of surrounding vehicles is a key competence for safe and efficient real-world autonomous driving systems. Previous works have presented deep neural network models for predictions using a detailed prior map which includes driving lanes and explicitly expresses the road rules like legal traffic directions and valid paths through intersections. Since it is unrealistic to assume the existence of the detailed prior maps for all areas, we use a map generated from only perceptual data (3D points measured by a LiDAR sensor). Such maps do not explicitly denote road rules, which makes prediction tasks more difficult. To overcome this problem, we propose a novel generative adversarial network (GAN) based framework. A discriminator in our framework can distinguish whether predicted trajectories follow road rules, and a generator can predict trajectories following it. Our framework implicitly extracts road rules by projecting trajectories onto the map via a differentiable function and training positional relations between trajectories and obstacles on the map. We also extend our framework to multimodal predictions so that various future trajectories are predicted. Experimental results show that our method outperforms other state-of-the-art methods in terms of trajectory errors and the ratio of trajectories that fall on drivable lanes.

****

Handwritten Chinese Font Generation With Collaborative Stroke Refinement

Chuan Wen, Yujie Pan, Jie Chang, Ya Zhang, Siheng Chen, Yanfeng Wang, Mei Han, Qi Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3882-3891

Automatic character generation is an appealing solution for typeface design, especially for Chinese fonts with over 3700 most commonly-used characters. This task is particularly challenging for handwritten characters with thin strokes which are error-prone during deformation. To handle the generation of thin strokes, w e introduce an auxiliary branch for stroke refinement. The auxiliary branch is trained to generate the bold version of target characters which are then fed to the dominating branch to guide the stroke refinement. The two branches are jointly trained in a collaborative fashion. In addition, for practical use, it is desirable to train the character synthesis model with a small set of manually designed characters. Taking advantage of content-reuse phenomenon in Chinese characters, we further propose an online zoom-augmentation strategy to reduce the dependency on large size training sets. The proposed model is trained end-to-end and can be added on top of any method for font synthesis. Experimental results on handwritten font synthesis have shown that the proposed method significantly outperforms the state-of-the-art methods under practical setting, i.e. with only 750 paired training samples.

****

Ontology-Driven Event Type Classification in Images

Eric Muller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, Ralph Ewerth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2928-2938

Event classification can add valuable information for semantic search and the increasingly important topic of fact validation in news. So far, only few approaches address image classification for newsworthy event types such as natural disasters, sports events, or elections. Previous work distinguishes only between a limited number of event types and relies on rather small datasets for training. In this paper, we present a novel ontology-driven approach for the classification of event types in images. We leverage a large number of real-world news events to pursue two objectives: First, we create an ontology based on Wikidata comprising the majority of event types. Second, we introduce a novel large-scale dataset that was acquired through Web crawling. Several baselines are proposed including an ontology-driven learning approach that aims to exploit structured information of a knowledge graph to learn relevant event relations using deep neural networks. Experimental results on existing as well as novel benchmark datasets demonstrate the superiority of the proposed ontology-driven approach.

****

Multimodal Prototypical Networks for Few-Shot Learning

Frederik Pahde, Mihai Puscas, Tassilo Klein, Moin Nabi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2644-2653

Although providing exceptional results for many computer vision tasks, state-of-the-art deep learning algorithms catastrophically struggle in low data scenarios. However, if data in additional modalities exist (e.g. text) this can compensate for the lack of data and improve the classification results. To overcome this data scarcity, we design a cross-modal feature generation framework capable of enriching the low populated embedding space in few-shot scenarios, leveraging data from the auxiliary modality. Specifically, we train a generative model that maps text data into the visual feature space to obtain more reliable prototypes. This allows to exploit data from additional modalities (e.g. text) during training while the ultimate task at test time remains classification with exclusively visual data. We show that in such cases nearest neighbor classification is a viable approach and outperform state-of-the-art single-modal and multimodal few-shot learning methods on the CUB-200 and Oxford-102 datasets.

****

Temporally Consistent 3D Human Pose Estimation Using Dual 360deg Cameras

Matthew Shere, Hansung Kim, Adrian Hilton; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 81-90

This paper presents a 3D human pose estimation system that uses a stereo pair of 360deg sensors to capture the complete scene from a single location. The approach combines the advantages of omnidirectional capture, the accuracy of multiple view 3D pose estimation and the portability of monocular acquisition. Joint monocular belief maps for joint locations are estimated from 360deg images and are used to fit a 3D skeleton to each frame. Temporal data association and smoothing is performed to produce accurate 3D pose estimates throughout the sequence. We e valuate our system on the Panoptic Studio dataset, as well as real 360deg video for tracking multiple people, demonstrating an average Mean Per Joint Position Error of 124.73mm with 30cm baseline cameras. We also demonstrate improved capabilities over perspective and 360deg multi-view systems when presented with limited camera views of the subject.

****

The IKEA ASM Dataset: Understanding People Assembling Furniture Through Actions, Objects and Pose

Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 847-859

The availability of a large labelled dataset is a key requirement for applying deep learning methods to solve various computer vision tasks. In the context of understanding human activities, existing public datasets, while large in size, are often limited to a single RGB camera and provide only per-frame or per-clip action annotations. To enable richer analysis and understanding of human activities, we introduce IKEA ASM---a three million frame, multi-view, furniture assembly video dataset that includes depth, atomic actions, object segmentation, and human poses. Additionally, we benchmark prominent methods for video action recognition, object segmentation and human pose estimation tasks on this challenging dataset. The dataset enables the development of holistic methods, which integrate multi-modal and multi-view data to better perform on these tasks.

****

Vid2Int: Detecting Implicit Intention From Long Dialog Videos

Xiaoli Xu, Yao Lu, Zhiwu Lu, Tao Xiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3299-3308

Detecting subtle intention such as deception and subtext of a person in a long dialog video, or implicit intention detection (IID), is a challenging problem. The transcript (textual cues) often reveals little, so audio-visual cues including voice tone as well as facial and body behaviour are the main focuses for automated IID. Contextual cues are also crucial, since a person's implicit intentions are often correlated and context-dependent when the person moves from one question-answer pair to the next. However, no such dataset exists which contains fine-grained question-answer pair (video segment) level annotation. The first contribution of this work is thus a new benchmark dataset, called Vid2Int-Deception to fill this gap. A novel multi-grain representation model is also proposed to capture the subtle movement changes of eyes, face, and body (relevant for inferring intention) from a long dialog video. Moreover, to model the temporal correlation between the implicit intentions across video segments, we propose a Video-to-Intention network (Vid2Int) based on attentive recurrent neural network (RNN). Extensive experiments show that our model achieves state-of-the-art results.

****

This Face Does Not Exist... But It Might Be Yours! Identity Leakage in Generative Models

Patrick Tinsley, Adam Czajka, Patrick Flynn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1320-1328

Generative adversarial networks (GANs) are able to generate high resolution photo-realistic images of objects that "do not exist." These synthetic images are rather difficult to detect as fake. However, the manner in which these generative models are trained hints at a potential for information leakage from the supplied training data, especially in the context of synthetic faces. This paper presents experiments suggesting that identity information in face images can flow from the training corpus into synthetic samples without any adversarial actions when building or using the existing model. This raises privacy-related questions, but also stimulates discussions of (a) the face manifold's characteristics in the feature space and (b) how to create generative models that do not inadvertently reveal identity information of real subjects whose images were used for training. We used five different face matchers (face_recognition, FaceNet, ArcFace, SphereFace and Neurotechnology MegaMatcher) and the StyleGAN2 synthesis model, and show that this identity leakage does exist for some, but not all methods. So, can

we say that these synthetically generated faces truly do not exist? Databases of real and synthetically generated faces are made available with this paper to allow full replicability of the results discussed in this work.
****

Adversarial Dual Distinct Classifiers for Unsupervised Domain Adaptation
Taotao Jing, Zhengming Ding; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 605-614

Unsupervised Domain adaptation (UDA) attempts to recognize the unlabeled target samples by building a learning model from a differently-distributed labeled source domain. Conventional UDA concentrates on extracting domain-invariant features through deep adversarial networks. However, most of them seek to match the different domain feature distributions, without considering the task-specific decision boundaries across various classes. In this paper, we propose a novel Adversarial Dual Distinct Classifiers Network (AD^2CN) to align the source and target domain data distribution simultaneously with matching task-specific category boundaries. To be specific, a domain-invariant feature generator is exploited to embed the source and target data into a latent common space with the guidance of discriminative cross-domain alignment. Moreover, we naturally design two different structure classifiers to identify the unlabeled target samples over the supervision of the labeled source domain data. Such dual distinct classifiers with various architectures can capture diverse knowledge of the target data structure from different perspectives. Extensive experimental results on several cross-domain visual benchmarks prove the model's effectiveness by comparing it with other state-of-the-art UDA.
****

Improving Point Cloud Semantic Segmentation by Learning 3D Object Detection
Ozan Unal, Luc Van Gool, Dengxin Dai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2950-2959

Point cloud semantic segmentation plays an essential role in autonomous driving, providing vital information about drivable surfaces and nearby objects that can aid higher level tasks such as path planning and collision avoidance. While current 3D semantic segmentation networks focus on convolutional architectures that perform great for well represented classes, they show a significant drop in performance for underrepresented classes that share similar geometric features. We propose a novel Detection Aware 3D Semantic Segmentation (DASS) framework that explicitly leverages localization features from an auxiliary 3D object detection task. By utilizing multitask training, the shared feature representation of the network is guided to be aware of per class detection features that aid tackling the differentiation of geometrically similar classes. We additionally provide a pipeline that uses DASS to generate high recall proposals for existing 2-stage detectors and demonstrate that the added supervisory signal can be used to improve 3D orientation estimation capabilities. Extensive experiments on both the SemanticKITTI and KITTI object datasets show that DASS can improve 3D semantic segmentation results of geometrically similar classes up to 37.8% IoU in image FOV while maintaining high precision bird's-eye view (BEV) detection results.
****

Optimistic Agent: Accurate Graph-Based Value Estimation for More Successful Visual Navigation
Mahdi Kazemi Moghaddam, Qi Wu, Ehsan Abbasnejad, Javen Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3733-3742

We humans can impeccably search for a target object, given its name only, even in an unseen environment. We argue this ability is largely due to three main reasons: the incorporation of prior knowledge (or experience), the adaptation of it to the new environment using the observed visual cues and most importantly optimistically searching without giving up early.This is currently missing in the state-of-the-art visual navigation methods based on Reinforcement Learning (RL). In this paper, we propose to use externally learned prior knowledge of the relative object locations and integrate it into our model by constructing a neural graph. In order to efficiently incorporate the graph without increasing the state-sp ace complexity, we propose Graph-based Value Estimation (GVE) module. GVE provides a more accurate baseline for estimating the Advantage function in actor-critic RL algorithm. This results in reduced value estimation error and, consequently, convergence to a more optimal policy. Through empirical studies, we show that our agent, dubbed as the optimistic agent, has a more realistic estimate of the state value during a navigation episode which leads to a higher success rate. Our extensive ablation studies show the efficacy of our simple method which achieves the state-of-the-art results measured by the conventional visual navigation metrics, e.g. Success Rate (SR) and Success weighted by Path Length (SPL), in AI2THOR environment.

****

Local to Global: Efficient Visual Localization for a Monocular Camera

Sang Jun Lee, Deokhwa Kim, Sung Soo Hwang, Donghwan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2231-2240

Robust and accurate visual localization is one of the most fundamental elements in various technologies, such as autonomous driving and augmented reality. While recent visual localization algorithms demonstrate promising results in terms of accuracy and robustness, the associated high computational cost requires running these algorithms on server-sides rather than client devices. This paper proposes a real time monocular visual localization system that combines client-side visual odometry with server-side visual localization functionality. In particular, the proposed system utilizes handcrafted features for real time visual odometry while adopting learned features for robust visual localization. To link the two components, the proposed system employs a map alignment mechanism that transforms the local coordinates obtained using visual odometry to global coordinates. The system achieves comparable accuracy to that of the state-of-the-art structure-based methods and end-to-end methods for the visual localization on both indoor and outdoor datasets while operating in real time.

****

The Laughing Machine: Predicting Humor in Video

Yuta Kayatani, Zekun Yang, Mayu Otani, Noa Garcia, Chenhui Chu, Yuta Nakashima, Haruo Takemura; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2073-2082

Humor is a very important communication tool; yet, it is an open problem for machines to understand humor. In this paper, we build a new multimodal dataset for humor prediction that includes subtitles and video frames, as well as humor labels associated with video's timestamps. On top of it, we present a model to predict whether a subtitle causes laughter. Our model uses the visual modality through facial expression and character name recognition, together with the verbal modality, to explore how the visual modality helps. In addition, we use an attention mechanism to adjust the weight for each modality to facilitate humor prediction. Interestingly, our experimental results show that the performance boost by combinations of different modalities, and the attention mechanism and the model mostly relies on the verbal modality.

****

Transductive Visual Verb Sense Disambiguation

Sebastiano Vascon, Sinem Aslan, Gianluca Bigaglia, Lorenzo Giudice, Marcello Pelillo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3050-3059

Verb Sense Disambiguation is a well-known task in NLP, the aim is to find the correct sense of a verb in a sentence. Recently, this problem has been extended in a multimodal scenario, by exploiting both textual and visual features of ambiguous verbs leading to a new problem, the Visual Verb Sense Disambiguation (VVSD). Here, the sense of a verb is assigned considering the content of an image paired with it rather than a sentence in which the verb appears. Annotating a dataset for this task is more complex than textual disambiguation, because assigning the correct sense to a pair of <image, verb> requires both non-trivial linguistic and visual skills. In this work, differently from the literature, the VVSD task will be performed in a transductive semi-supervised learning (SSL) setting, in w

hich only a small amount of labeled information is required, reducing tremendously the need for annotated data. The disambiguation process is based on a graph-based label propagation method which takes into account mono or multimodal representations for <image, verb> pairs. Experiments have been carried out on the recently published dataset VerSe, the only available dataset for this task. The achieved results outperform the current state-of-the-art by a large margin while using only a small fraction of labeled samples per sense. The code is available: https://github.com/GiBg1aN/TVVSD.
****
Multi-Loss Weighting With Coefficient of Variations
Rick Groenendijk, Sezer Karaoglu, Theo Gevers, Thomas Mensink; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1469-1478
Many interesting tasks in machine learning and computer vision are learned by optimising an objective function defined as a weighted linear combination of multiple losses. The final performance is sensitive to choosing the correct (relative) weights for these losses. Finding a good set of weights is often done by adopting them into the set of hyper-parameters, which are set using an extensive grid search. This is computationally expensive. In this paper, we propose a weighting scheme based on the coefficient of variations and set the weights based on properties observed while training the model. The proposed method incorporates a measure of uncertainty to balance the losses, and as a result the loss weights evolve during training without requiring another (learning based) optimisation. In contrast to many loss weighting methods in literature, we focus on single-task multi-loss problems, such as monocular depth estimation and semantic segmentation, and show that multi-task approaches for loss weighting do not work on those single-tasks. The validity of the approach is shown empirically for depth estimation and semantic segmentation on multiple datasets.
****
De-Biasing Neural Networks With Estimated Offset for Class Imbalanced Learning
Byungju Kim, Hyeong Gwon Hong, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1479-1487
The imbalanced distribution of the training data makes the networks biased to the frequent classes. Existing methods to resolve the problem involve re-sampling, re-weighting, or cost-sensitive learning. Most of them anticipate that emphasizing the minority classes during the training would help the network to learn better representations. In this paper, we propose a method for reparameterizing softmax classifiers' offsets so that training is less sensitive to class imbalance. We first observe that the trained offset of the baseline linear classifier is biased toward the majority classes due to the imbalance. Instead of the trained offset, we define the estimated offset, and constrain it to be uniform over the classes. In experiments with long-tailed benchmarks, our method exhibits the best performance. These experiments verify that our proposed method effectively encourages the networks to learn better representations for minority classes while preserving the performance for the majority classes.
****
Oriented Object Detection in Aerial Images With Box Boundary-Aware Vectors
Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, Dimitris Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2150-2159
Oriented object detection in aerial images is a challenging task as the objects in aerial images are displayed in arbitrary directions and are usually densely packed. Current oriented object detection methods mainly rely on two-stage anchor-based detectors. However, the anchor-based detectors typically suffer from a severe imbalance issue between the positive and negative anchor boxes. To address this issue, in this work we extend the horizontal keypoint-based object detector to the oriented object detection task. In particular, we first detect the center keypoints of the objects, based on which we then regress the box boundary-aware vectors (BBAVectors) to capture the oriented bounding boxes. The box boundary-aware vectors are distributed in the four quadrants of a Cartesian coordinate sy

stem for all arbitrarily oriented objects. To relieve the difficulty of learning the vectors in the corner cases, we further classify the oriented bounding boxes into horizontal and rotational bounding boxes. In the experiment, we show that learning the box boundary-aware vectors is superior to directly predicting the width, height, and angle of an oriented bounding box, as adopted in the baseline method. Besides, the proposed method competes favorably with state-of-the-art methods. Code is available at https://github.com/yijingru/BBAVectors-Oriented-Object-Detection.

****

IncreACO: Incrementally Learned Automatic Check-Out With Photorealistic Exemplar Augmentation

Yandan Yang, Lu Sheng, Xiaolong Jiang, Haochen Wang, Dong Xu, Xianbin Cao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 626-634

Automatic check-out (ACO) emerges as an integral component in recent self-service retailing stores, which aims at automatically detecting and counting the randomly placed products upon a check-out platform. Existing data-driven counting works still have difficulties in generalizing to real-world retail product counting scenarios, since (1) real check-out images are hard to collect or cover all products and their possible layouts, (2) rapid updating of the product list leads to frequent and tedious re-training of the counting models. To overcome these obstacles, we contribute a practical automatic check-out framework tailored to real-world retail product counting scenarios, consisting of a photorealistic exemplar augmentation to generate physically reliable and photorealistic check-out images from canonical exemplars scanned for each product, and an incremental learning strategy to match the updating nature of the ACO system with much fewer training effort. Through comprehensive studies, we show that the proposed IncreACO serves as an effective framework on recent Retail Product Checkout (RPC) dataset, where the proposed photorealistic exemplar augmentation remarkably improves the counting performance against the state-of-the-art methods (77.15% v.s. 72.83% in counting accuracy), whilst the proposed incremental learning framework consistently extends the counting performance to new categories.

****

Understanding the Impact of Mistakes on Background Regions in Crowd Counting

Davide Modolo, Bing Shuai, Rahul Rama Varior, Joseph Tighe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1650-1659

In crowd counting we often observe wrong predictions on image regions not containing any person. But how often do these mistakes happen and how much do they affect the overall performance? In this paper we analyze this problem in depth and present an extensive analysis on five of the most important crowd counting datasets. We present this analysis in two parts. First, we quantify the number of mistakes made. Our results show that (i) mistakes on background are substantial and they are responsible for 18-49% of the total error, (ii) models do not generalize well to different kinds of backgrounds and perform poorly on completely background images, and (iii) models make many more mistakes than those captured by the standard Mean Absolute Error (MAE) metric, as counting on background compensates substantially for misses on foreground. And second, we quantify the performance change gained by helping the model better deal with this problem. We enrich a popular crowd counting network with a segmentation branch trained to suppress background predictions. This simple addition (i) reduces background error by 10-83%, (ii) reduces foreground error by up to 26% and (iii) improves overall crowd counting performance up to 20%. When compared against the literature, this simple technique achieves very competitive results on all datasets, showing the importance of tackling the background problem.

****

Temporal-Aware Self-Supervised Learning for 3D Hand Pose and Mesh Estimation in Videos

Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 202

1, pp. 1050-1059

Estimating 3D hand pose directly from RGB images is challenging but has gained steady progress recently by training deep models with annotated 3D poses. However annotating 3D poses is difficult and as such only a few 3D hand pose datasets are available, all with limited sample sizes. In this study, we propose a new framework of training 3D pose estimation models from RGB images without using explicit 3D annotations, i.e., trained with only 2D information. Our framework is motivated by two observations: 1) Videos provide richer information for estimating 3D poses as opposed to static images; 2) Estimated 3D poses ought to be consistent whether the videos are viewed in the forward order or reverse order. We leverage these two observations to develop a self-supervised learning model called temporal-aware self-supervised network (TASSN).By enforcing temporal consistency constraints, TASSN learns 3D hand poses and meshes from videos with only 2D keypoint position annotations. Experiments show that our model achieves surprisingly good results, with 3D estimation accuracy on par with the state-of-the-art models trained with 3D annotations, highlighting the benefit of the temporal consistency in constraining 3D prediction models.

****

Saliency Prediction With External Knowledge

Yifeng Zhang, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 484-493

The last decades have seen great progress in saliency prediction, with the success of deep neural networks that are able to encode high-level semantics. Yet, while humans have the innate capability in leveraging their knowledge to decide where to look (e.g. people pay more attention to familiar faces such as celebrities), saliency prediction models have only been trained with large eye-tracking datasets. This work proposes to bridge this gap by explicitly incorporating external knowledge for saliency models as humans do. We develop networks that learn to highlight regions by incorporating prior knowledge of semantic relationships, be it general or domain-specific, depending on the task of interest. At the core of the method is a new GraSSNet that constructs a graph that encodes semantic relationships learned from external knowledge. A Spatial Graph Attention Network is then developed to update saliency features based on the learned graph. Experiments show that the proposed model learns to predict saliency from the external knowledge and outperforms the state-of-the-art on four saliency benchmarks.

****

SuPEr-SAM: Using the Supervision Signal From a Pose Estimator to Train a Spatial Attention Module for Personal Protective Equipment Recognition

Adrian Sandru, Georgian-Emilian Duta, Mariana-Iuliana Georgescu, Radu Tudor Ionescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2817-2826

We propose a deep learning method to automatically detect personal protective equipment (PPE), such as helmets, surgical masks, reflective vests, boots and so on, in images of people. Typical approaches for PPE detection based on deep learning are (i) to train an object detector for items such as those listed above or (ii) to train a person detector and a classifier that takes the bounding boxes predicted by the detector and discriminates between people wearing and people not wearing the corresponding PPE items. We propose a novel and accurate approach that uses three components: a person detector, a body pose estimator and a classifier. Our novelty consists in using the pose estimator only at training time, to improve the prediction performance of the classifier. We modify the neural architecture of the classifier by adding a spatial attention mechanism, which is trained using supervision signal from the pose estimator. In this way, the classifier learns to focus on PPE items, using knowledge from the pose estimator with almost no computational overhead during inference.

****

A Unified Framework for Compressive Video Recovery From Coded Exposure Techniques

Prasan Shedligeri, Anupama S, Kaushik Mitra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1600-1609

Several coded exposure techniques have been proposed for acquiring high frame rate videos at low bandwidth. Most recently, a Coded-2-Bucket camera has been proposed that can acquire two compressed measurements in a single exposure, unlike previously proposed coded exposure techniques, which can acquire only a single measurement. Although two measurements are better than one for an effective video recovery, we are yet unaware of the clear advantage of two measurements, either quantitatively or qualitatively. Here, we propose a unified learning-based framework to make such a qualitative and quantitative comparison between those which capture only a single coded image (Flutter Shutter, Pixel-wise coded exposure) and those that capture two measurements per exposure (C2B). Our learning-based framework consists of a shift-variant convolutional layer followed by a fully convolutional deep neural network. Our proposed unified framework achieves the state of the art reconstructions in all three sensing techniques. Further analysis shows that when most scene points are static, the C2B sensor has a significant advantage over acquiring a single pixel-wise coded measurement. However, when most scene points undergo motion, the C2B sensor has only a marginal benefit over the single pixel-wise coded exposure measurement.

****

Deep Unsupervised Anomaly Detection

Tangqing Li, Zheng Wang, Siying Liu, Wen-Yan Lin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3636-3645

This paper proposes a novel method to detect anomalies in large datasets under a fully unsupervised setting. The key idea behind our algorithm is to learn the representation underlying normal data. To this end, we leverage the latest clustering technique suitable for handling high dimensional data. This hypothesis provides a reliable starting point for normal data selection. We train an autoencoder from the normal data subset, and iterate between hypothesizing nor-mal candidate subset based on clustering and representation learning. The reconstruction error from the learned autoencoder serves as a scoring function to assess the normality of the data. Experimental results on several public benchmark datasets show that the proposed method outperforms state-of-the-art unsupervised techniques and is comparable to semi-supervised techniques in most cases

****

Disentangled Contour Learning for Quadrilateral Text Detection

Yanguang Bi, Zhiqiang Hu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 909-918

Precise detection of quadrilateral text is of great significance for subsequent recognition, where the main challenge comes from four distorted sides. Existing methods concentrate on learning four vertices to construct the contour. However, vertices are dummy intersections entangled by their neighbor sides. The regression of each vertex would simultaneously affect its two neighbor sides. As a result, the originally independent side would be influenced by two different vertices which further inevitably disturb other sides. The above entangled vertices learning suppresses the learning efficiency and detection performance. In this paper, we proposed disentangled contour learning network (DCLNet) to focus on clear regression of each individual side disentangled from the whole quadrilateral contour. The side is parameterized by a linear equation that disentangled in the polar coordinates for easier learning. With tailored Ray-IoU loss and sine angle loss, DCLNet could better learn the representation of each disentangled side without being disturbed by others. The final quadrilateral text contour is easily constructed by intersecting the predicted linear equations of sides. Empirically, the proposed DCLNet achieves state-of-the-art detection performances on three scene text benchmarks. Ablation study is also presented to demonstrate the effectiveness of proposed disentangled contour learning framework.

****

DynaVSR: Dynamic Adaptive Blind Video Super-Resolution

Suyoung Lee, Myungsub Choi, Kyoung Mu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2093-2102

Most conventional supervised super-resolution (SR) algorithms assume that low-resolution (LR) data is obtained by downscaling high-resolution (HR) data with a f ixed known kernel, but such an assumption often does not hold in real scenarios. Some recent blind SR algorithms have been proposed to estimate different downscaling kernels for each input LR image. However, they suffer from heavy computational overhead, making them infeasible for direct application to videos. In this work, we present DynaVSR, a novel meta-learning-based framework for real-world video SR that enables efficient downscaling model estimation and adaptation to the current input. Specifically, we train a multi-frame downscaling module with various types of synthetic blur kernels, which is seamlessly combined with a video SR network for input-aware adaptation. Experimental results show that DynaVSR consistently improves the performance of the state-of-the-art video SR models by a large margin, with an order of magnitude faster inference time compared to the existing blind SR approaches.

****

MEVA: A Large-Scale Multiview, Multimodal Video Dataset for Activity Detection

Kellie Corona, Katie Osterdahl, Roderic Collins, Anthony Hoogs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1060-1068

We present the Multiview Extended Video with Activities (MEVA) dataset, a new and very-large-scale dataset for human activity recognition. Existing security datasets either focus on activity counts by aggregating public video disseminated due to its content, which typically excludes same-scene background video, or they achieve persistence by observing public areas and thus cannot control for activity content. Our dataset is over 9300 hours of untrimmed, continuous video, scripted to include diverse, simultaneous activities, along with spontaneous background activity. We have annotated 144 hours for 37 activity types, marking bounding boxes of actors and props. Our collection observed approximately 100 actors performing scripted scenarios and spontaneous background activity over a three-week period at access-controled venue, collecting in multiple modalities with overlapping and non-overlapping indoor and outdoor viewpoints. The resulting data includes video from 38 RGB and thermal IR cameras, 42 hours of UAV footage, as well as GPS locations for the actors. 122 hours of annotation are sequestered in support of the NIST Activity in Extended Video (ActEV) challenge; the other 22 hours of annotation and the corresponding video are available on our website, along with an additional 306 hours of ground camera data, 4.6 hours of UAV data, and 9.6 hours of GPS logs. Additional derived data includes camera models geo-registering the outdoor cameras and a dense 3D point cloud model of the outdoor scene. The data was collected with IRB oversight and approval and released under a CC-BY-4.0 license.

****

Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples

Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, Julian McAuley; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3348-3357

Recent advances in video manipulation techniques have made the generation of fake videos more accessible than ever before. Manipulated videos can fuel disinformation and reduce trust in media. Therefore detection of fake videos has garnered immense interest in academia and industry. Recently developed Deepfake detection methods rely on deep neural networks (DNNs) to distinguish AI-generated fake videos from real videos. In this work, we demonstrate that it is possible to bypass such detectors by adversarially modifying fake videos synthesized using existing Deepfake generation methods. We further demonstrate that our adversarial perturbations are robust to image and video compression codecs, making them a real-world threat. We present pipelines in both white-box and black-box attack scenarios that can fool DNN based Deepfake detectors into classifying fake videos as real.

****

PNPDet: Efficient Few-Shot Detection Without Forgetting via Plug-and-Play Sub-Networks

Gongjie Zhang, Kaiwen Cui, Rongliang Wu, Shijian Lu, Yonghong Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3823-3832

The human visual system can detect objects of unseen categories from merely a few examples. However, such capability remains absent in state-of-the-art detectors. To bridge this gap, several attempts have been proposed to perform few-shot detection by incorporating meta-learning techniques. Such methods can improve detection performance on unseen categories, but also add huge computational burden, and usually degrade detection performance on seen categories. In this paper, we present PNPDet, a novel Plug-and-Play Detector, for efficient few-shot detection without forgetting. It introduces a simple but effective architecture with separate sub-networks that disentangles the recognition of base and novel categories and prevents hurting performance on known categories while learning new concepts. Distance metric learning is further incorporated into sub-networks, consistently boosting detection performance for both base and novel categories. Experiments show that the proposed PNPDet can achieve comparable few-shot detection performance on unseen categories while not losing accuracy on seen categories, and also remain efficient and flexible at the same time.

****

Conflicting Bundles: Adapting Architectures Towards the Improved Training of Deep Neural Networks

David Peer, Sebastian Stabinger, Antonio Rodriguez-Sanchez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 256-265

Designing neural network architectures is a challenging task and knowing which specific layers of a model must be adapted to improve the performance is almost a mystery. In this paper, we introduce a novel theory and metric to identify layers that decrease the test accuracy of the trained models, this identification is done as early as at the beginning of training. In the worst-case, such a layer could lead to a network that can not be trained at all. More precisely, we identified those layers that worsen the performance because they produce conflicting training bundles as we show in our novel theoretical analysis, complemented by our extensive empirical studies. Based on these findings, a novel algorithm is introduced to remove performance decreasing layers automatically. Architectures found by this algorithm achieve a competitive accuracy when compared against the state-of-the-art architectures. While keeping such high accuracy, our approach drastically reduces memory consumption and inference time for different computer vision tasks.

****

Multi Projection Fusion for Real-Time Semantic Segmentation of 3D LiDAR Point Clouds

Yara Ali Alnaggar, Mohamed Afifi, Karim Amer, Mohamed ElHelw; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1800-1809

Semantic segmentation of 3D point cloud data is essential for enhanced high-level perception in autonomous platforms. Furthermore, given the increasing deployment of LiDAR sensors onboard of cars and drones, a special emphasis is also placed on non-computationally intensive algorithms that operate on mobile GPUs. Previous efficient state-of-the-art methods relied on 2D spherical projection of point clouds as input for 2D fully convolutional neural networks to balance the accuracy-speed trade-off. This paper introduces a novel approach for 3D point cloud semantic segmentation that exploits multiple projections of the point cloud to mitigate the loss of information inherent in single projection methods. Our Multi-Projection Fusion (MPF) framework analyzes spherical and bird's-eye view projections using two separate highly-efficient 2D fully convolutional models then combines the segmentation results of both views. The proposed framework is validated on the SemanticKITTI dataset where it achieved a mIoU of 55.5 which is higher than state-of-the-art projection-based methods RangeNet++ [23] and PolarNet [44] while being 1.6x faster than the former and 3.1x faster than the latter.

****

Cinematic-L1 Video Stabilization With a Log-Homography Model

Arwen Bradley, Jason Klivington, Joseph Triscari, Rudolph van der Merwe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1041-1049

We present a method for stabilizing handheld video that simulates the camera motions cinematographers achieve with equipment like tripods, dollies, and Steadicams. We formulate a constrained convex optimization problem minimizing the L1-norm of the first three derivatives of the stabilized motion. Our approach extends the work of Grundmann et al. [9] by solving with full homographies (rather than affinities) in order to correct perspective, preserving linearity by working in log-homography space. We also construct crop constraints that preserve field-of-view; model the problem as a quadratic (rather than linear) program to allow for an L2 term encouraging fidelity to the original trajectory; and add constraints and objectives to reduce distortion. Furthermore, we propose new methods for handling salient objects via both inclusion constraints and centering objectives. Finally, we describe a windowing strategy to approximate the solution in linear time and bounded memory. Our method is computationally efficient, running at 300 fps on an iPhone XS, and yields high-quality results, as we demonstrate with a collection of stabilized videos, quantitative and qualitative comparisons to [9] and other methods, and an ablation study.

****

Temporal Context Aggregation for Video Retrieval With Contrastive Learning

Jie Shao, Xin Wen, Bingchen Zhao, Xiangyang Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3268-3278

The current research focus on Content-Based Video Retrieval requires higher-level video representation describing the long-range semantic dependencies of relevant incidents, events, etc. However, existing methods commonly process the frames of a video as individual images or short clips, making the modeling of long-range semantic dependencies difficult. In this paper, we propose TCA (Temporal Context Aggregation for Video Retrieval), a video representation learning network that incorporates long-range temporal information between frame-level features using the self-attention mechanism for video retrieval. To train it on video retrieval datasets, we propose a supervised contrastive learning method that performs automatic hard negative mining and utilizes the memory bank mechanism to increase the capacity of negative samples. Extensive experiments are conducted on multiple video retrieval tasks, such as CC_WEB_VIDEO, FIVR-200K, and EVVE. The proposed method shows a significant performance advantage (17% mAP on FIVR-200K) over state-of-the-art methods with video-level features, and deliver competitive results with 22x faster inference time comparing with frame-level features.

****

TrustMAE: A Noise-Resilient Defect Classification Framework Using Memory-Augmented Auto-Encoders With Trust Regions

Daniel Stanley Tan, Yi-Chun Chen, Trista Pei-Chun Chen, Wei-Chao Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 276-285

In this paper, we propose a framework called TrustMAE to address the problem of product defect classification. Instead of relying on defective images that are difficult to collect and laborious to label, our framework can accept datasets with unlabeled images. Moreover, unlike most anomaly detection methods, our approach is robust against noises, or defective images, in the training dataset. Our framework uses a memory-augmented auto-encoder with a sparse memory addressing scheme to avoid over-generalizing the auto-encoder, and a novel trust-region memory updating scheme to keep the noises away from the memory slots. The result is a framework that can reconstruct defect-free images and identify the defective regions using a perceptual distance network. When compared against various state-of-the-art baselines, our approach performs competitively under noise-free MVTec datasets. More importantly, it remains effective at a noise level up to 40% while significantly outperforming other baselines.

****

Visual Speech Enhancement Without a Real Visual Stream

Sindhu B. Hegde, K.R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V.

Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1926-1935

In this work, we re-think the task of speech enhancement in unconstrained real-world environments. Current state-of-the-art methods use only the audio stream and are limited in their performance in a wide range of real-world noises. Recent works using lip movements as additional cues improve the quality of generated speech over "audio-only" methods. But, these methods cannot be used for several applications where the visual stream is unreliable or completely absent. We propose a new paradigm for speech enhancement by exploiting recent breakthroughs in speech-driven lip synthesis. Using one such model as a teacher network, we train a robust student network to produce accurate lip movements that mask away the noise, thus acting as a "visual noise filter". The intelligibility of the speech enhanced by our pseudo-lip approach is comparable (< 3% difference) to the case of using real lips. This implies that we can exploit the advantages of using lip movements even in the absence of a real video stream. We rigorously evaluate our model using quantitative metrics as well as human evaluations. Additional ablation studies and a demo video on our website containing qualitative comparisons and results clearly illustrate the effectiveness of our approach.

****

Dynamic Routing Networks

Shaofeng Cai, Yao Shu, Wei Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3588-3597

The deployment of deep neural networks in real-world applications is mostly restricted by their high inference costs. Extensive efforts have been made to improve the accuracy with expert-designed or algorithm-searched architectures. However, the incremental improvement is typically achieved with increasingly more expensive models that only a small portion of input instances really need. Inference with a static architecture that processes all input instances via the same transformation would thus incur unnecessary computational costs. Therefore, customizing the model capacity in an instance-aware manner is much needed for higher inference efficiency. In this paper, we propose Dynamic Routing Networks (DRNets), which support efficient instance-aware inference by routing the input instance to only necessary transformation branches selected from a candidate set of branches for each connection between transformation nodes. The branch selection is dynamically determined via the corresponding branch importance weights, which are first generated from lightweight hypernetworks (RouterNets) and then recalibrated with Gumbel-Softmax before the selection. Extensive experiments show that DRNets can reduce a substantial amount of parameter size and FLOPs during inference with prediction performance comparable to state-of-the-art architectures.

****

Foreground Color Prediction Through Inverse Compositing

Sebastian Lutz, Aljosa Smolic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1610-1619

In natural image matting, the goal is to estimate the opacity of the foreground object in the image. This opacity controls the way the foreground and background is blended in transparent regions. In recent years, advances in deep learning have led to many natural image matting algorithms that have achieved outstanding performance in a fully automatic manner. However, most of these algorithms only predict the alpha matte from the image, which is not sufficient to create high-quality compositions. Further, it is not possible to manually interact with these algorithms in any way except by directly changing their input or output. We propose a novel recurrent neural network that can be used as a post-processing method to recover the foreground and background colors of an image, given an initial alpha estimation. Our method outperforms the state-of-the-art in color estimation for natural image matting and show that the recurrent nature of our method allows users to easily change candidate solutions that lead to superior color estimations.

****

Reducing the Annotation Effort for Video Object Segmentation Datasets

Paul Voigtlaender, Lishu Luo, Chun Yuan, Yong Jiang, Bastian Leibe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3060-3069

For further progress in video object segmentation (VOS), larger, more diverse, and more challenging datasets will be necessary. However, densely labeling every frame with pixel masks does not scale to large datasets. We use a deep convolutional network to automatically create pseudo-labels on a pixel level from much cheaper bounding box annotations and investigate how far such pseudo-labels can carry us for training state-of-the-art VOS approaches. A very encouraging result of our study is that adding a manually annotated mask in only a single video frame for each object is sufficient to generate pseudo-labels which can be used to train a VOS method to reach almost the same performance level as when training with fully segmented videos. We use this workflow to create pixel pseudo-labels for the training set of the challenging tracking dataset TAO, and we manually annotate a subset of the validation set. Together, we obtain the new TAO-VOS benchmark, which we make publicly available at www.vision.rwth-aachen.de/page/taovos. While the performance of state-of-the-art methods on existing datasets starts to saturate, TAO-VOS remains very challenging for current algorithms and reveals their shortcomings.

****

Style Consistent Image Generation for Nuclei Instance Segmentation

Xuan Gong, Shuyan Chen, Baochang Zhang, David Doermann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3994-4003

In medical image analysis, one limitation of the application of machine learning is the insufficient amount of data with detailed annotation, due primarily to high cost. Another impediment is the domain gap observed between images from different organs and different collections. The differences are even more challenging for the nuclei instance segmentation, where images have significant nuclei stain distribution variations and complex pleomorphisms (sizes and shapes). In this work, we generate style consistent histopathology images for nuclei instance segmentation. We set up a novel instance segmentation framework that integrates a generator and discriminator into the segmentation pipeline with adversarial training to generalize nuclei instances and texture patterns. A segmentation net detects and segments both real nuclei and synthetic nuclei and provides feedback so that the generator can synthesize images that can boost the segmentation performance. Experimental results on three public nuclei datasets indicate that our proposed method outperforms the state-of-the-art significantly.

****

FlowCaps: Optical Flow Estimation With Capsule Networks for Action Recognition

Vinoj Jayasundara, Debaditya Roy, Basura Fernando; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3409-3418

Capsule networks (CapsNets) have recently shown promise to excel in most computer vision tasks, especially pertaining to scene understanding. In this paper, we explore CapsNet's capabilities in optical flow estimation, a task at which convolutional neural networks (CNNs) have already outperformed other approaches. We propose a CapsNet-based architecture, termed FlowCaps, which attempts to a) achieve better correspondence matching via finer-grained, motion-specific, and more-interpretable encoding crucial for optical flow estimation, b) perform better-generalizable optical flow estimation, c) utilize lesser ground truth data, and d) significantly reduce the computational complexity in achieving good performance, in comparison to its CNN-counterparts.

****

Legacy Photo Editing With Learned Noise Prior

Yuzhi Zhao, Lai-Man Po, Tingyu Lin, Xuehui Wang, Kangcheng Liu, Yujia Zhang, Wing-Yin Yu, Pengfei Xian, Jingjing Xiong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2103-2112

There are quite a number of photographs captured under undesirable conditions in the last century. Thus, they are often noisy, regionally incomplete, and grayscale formatted. Conventional approaches mainly focus on one point so that those restoration results are not perceptually sharp or clean enough. To solve these pr oblems, we propose a noise prior learner NEGAN to simulate the noise distribution of real legacy photos using unpaired images. It mainly focuses on matching high-frequency parts of noisy images through discrete wavelet transform (DWT) since they include most of noise statistics. We also create a large legacy photo dataset for learning noise prior. Using learned noise prior, we can easily build valid training pairs by degrading clean images. Then, we propose an IEGAN framework performing image editing including joint denoising, inpainting and colorization based on the estimated noise prior. We evaluate the proposed system and compare it with state-of-the-art image enhancement methods. The experimental results demonstrate that it achieves the best perceptual quality. Please see the webpage https://github.com/zhaoyuzhi/Legacy-Photo-Editing-with-Learned-Noise-Prior for the codes and the proposed LP dataset.

****

AdarGCN: Adaptive Aggregation GCN for Few-Shot Learning

Jianhong Zhang, Manli Zhang, Zhiwu Lu, Tao Xiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3482-3491

Existing few-shot learning (FSL) methods assume that there exist sufficient training samples from source classes for knowledge transfer to target classes with few training samples. However, this assumption is often invalid, especially when it comes to fine-grained recognition. In this work, we define a new FSL setting termed scarce-source few-shot learning (SSFSL), under which both the source and target classes have limited training samples. To overcome the source class data scarcity problem, a natural option is to crawl images from the web with class names as search keywords. However, the crawled images are inevitably corrupted by large amount of noise (irrelevant images) and thus may harm the performance. To address this problem, we propose a graph convolutional network (GCN)-based label denoising (LDN) method to remove the irrelevant images. Further, with the cleaned web images as well as the original clean training images, we propose a GCN-based FSL method. For both the LDN and FSL tasks, a novel adaptive aggregation GCN (AdarGCN) model is proposed, which differs from existing GCN models in that adaptive aggregation is performed based on a multi-head multi-level aggregation module. With AdarGCN, how much and how far information carried by each graph node is propagated in the graph structure can be determined automatically, therefore alleviating the effects of both noisy and outlying training samples. Extensive experiments demonstrate the superior performance of our AdarGCN under both the new SSFSL and the conventional FSL settings.

****

Misclassification Risk and Uncertainty Quantification in Deep Classifiers

Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydogan, John Reid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2484-2492

In this paper, we propose risk-calibrated evidential deep classifiers to reduce the costs associated with classification errors. We use two main approaches. The first is to develop methods to quantify the uncertainty of a classifier's predictions and reduce the likelihood of acting on erroneous predictions. The second is a novel way to train the classifier such that erroneous classifications are biased towards less risky categories. We combine these two approaches in a principled way. While doing this, we extend evidential deep learning with pignistic probabilities, which are used to quantify uncertainty of classification predictions and model rational decision making under uncertainty. We evaluate the performance of our approach on several image classification tasks. We demonstrate that our approach allows to (i) incorporate misclassification cost while training deep classifiers, (ii) accurately quantify the uncertainty of classification predictions, and (iii) simultaneously learn how to make classification decisions to minimize expected cost of classification errors.

****

A Vector-Based Representation to Enhance Head Pose Estimation

Zhiwen Cao, Zongcheng Chu, Dongfang Liu, Yingjie Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1188-1197

This paper proposes to use the three vectors in a rotation matrix as the representation in head pose estimation and develops a new neural network based on the characteristic of such representation. We address two potential issues existed in current head pose estimation works: 1. Public datasets for head pose estimation use either Euler angles or quaternions to annotate data samples. However, both of these annotations have the issue of discontinuity and thus could result in some performance issues in neural network training. 2. Most research works report Mean Absolute Error (MAE) of Euler angles as the measurement of performance. We show that MAE may not reflect the actual behavior especially for the cases of profile views. To solve these two problems, we propose a new annotation method which uses three vectors to describe head poses and a new measurement Mean Absolute Error of Vectors (MAEV) to assess the performance. We also train a new neural network to predict the three vectors with the constraints of orthogonality. Our proposed method achieves state-of-the-art results on both AFLW2000 and BIWI datasets. Experiments show our vector-based annotation method can effectively reduce prediction errors for large pose angles.
****

Automatic Open-World Reliability Assessment
Mohsen Jafarzadeh, Touqeer Ahmad, Akshay Raj Dhamija, Chunchun Li, Steve Cruz, Terrance E. Boult; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1984-1993

Image classification in the open-world must handle out-of-distribution (OOD) images. Systems should ideally reject OOD images, or they will map atop of known classes and reduce reliability. Using open-set classifiers that can reject OOD inputs can help. However, optimal accuracy of open-set classifiers depend on the frequency of OOD data. Thus, for either standard or open-set classifiers, it is important to be able to determine when the world changes and increasing OOD inputs will result in reduced system reliability. However, during operations, we cannot directly assess accuracy as there are no labels. Thus, the reliability assessment of these classifiers must be done by human operators, made more complex because networks are not 100% accurate, so some failures are to be expected. To automate this process, herein, we formalize the open-world recognition reliability problem and propose multiple automatic reliability assessment policies to address this new problem using only the distribution of reported scores/probability data. The distributional algorithms can be applied to both classic classifiers with SoftMax as well as the open-world Extreme Value Machine (EVM) to provide automated reliability assessment. We show that all of the new algorithms significantly outperform detection using the mean of SoftMax.
****

SSGP: Sparse Spatial Guided Propagation for Robust and Generic Interpolation
Rene Schuster, Oliver Wasenmuller, Christian Unger, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 197-206

Interpolation of sparse pixel information towards a dense target resolution finds its application across multiple disciplines in computer vision. State-of-the-art interpolation of motion fields applies model-based interpolation that makes use of edge information extracted from the target image. For depth completion, data-driven learning approaches are widespread. Our work is inspired by latest trends in depth completion that tackle the problem of dense guidance for sparse information. We extend these ideas and create a generic cross-domain architecture that can be applied for a multitude of interpolation problems like optical flow, scene flow, or depth completion. In our experiments, we show that our proposed concept of Sparse Spatial Guided Propagation (SSGP) achieves improvements to robustness, accuracy, or speed compared to specialized algorithms.
****

Compositional Learning of Image-Text Query for Image Retrieval
Muhammad Umer Anwaar, Egor Labintcev, Martin Kleinsteuber; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1140-1149

In this paper, we investigate the problem of retrieving images from a database b ased on a multi-modal (image-text) query. Specifically, the query text prompts some modification in the query image and the task is to retrieve images with the desired modifications. For instance, a user of an e-commerce platform is interested in buying a dress, which should look similar to her friend's dress, but the dress should be of white color with a ribbon sash. In this case, we would like the algorithm to retrieve some dresses with desired modifications in the query dress. We propose an autoencoder based model, ComposeAE, to learn the composition of image and text query for retrieving images. We adopt a deep metric learning approach and learn a metric that pushes composition of source image and text query closer to the target images. We also propose a rotational symmetry constraint on the optimization problem. Our approach is able to outperform the state-of-the-art method TIRG on three benchmark datasets, namely: MIT-States, Fashion200k and Fashion IQ. In order to ensure fair comparison, we introduce strong baselines by enhancing TIRG method. To ensure reproducibility of the results, we publish our code here: https://github.com/ecom-research/ComposeAE.

****

Object Recognition With Continual Open Set Domain Adaptation for Home Robot

Ikki Kishida, Hong Chen, Masaki Baba, Jiren Jin, Ayako Amma, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1517-1526

Object recognition ability is indispensable for robots to act like humans in a home environment. For example, when considering an object searching task, humans can recognize a naturally arranged object previously held in their hands while ignoring never observed objects. Even in such a simple task, we need to deal with three complex problems: domain adaptation, open-set recognition, and continual learning. However, most existing datasets are simplified to focus on one problem and do not measure the object recognition ability for home robots when multiple problems are simultaneously present. In this paper, we propose the COSDA-HR (Continual Open Set Domain Adaptation for Home Robot) dataset that requires dealing with the above three problems simultaneously. The COSDA-HR dataset focuses particularly on the scenario in which naturally arranged objects in a room are recognized by training with handheld objects towards the goal of creating a user-friendly teaching system for home robots. We provide various baselines to address the problems in the COSDA-HR dataset by combining state-of-the-art methods from each research area and analyze the limitations of such simple combinations. We consider that it is necessary to study the methods of handling multiple problems simultaneously instead of solving each problem to realize practical object recognition systems for home robots.

****

MoRe: A Large-Scale Motorcycle Re-Identification Dataset

Augusto Figueiredo, Johnata Brayan, Renan Oliveira Reis, Raphael Prates, William Robson Schwartz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 4034-4043

Motorcycles are often related to transit and criminal issues due to its abundance in the transit. Despite its importance, motorcycles are a seldom addressed problem in the computer vision community. We credit this problem to the lack of large-scale datasets and strong baseline models. Therefore, we present the first large-scale Motorcycles Re-Identification (MoRe) dataset. MoRe consists of 3,827 individuals (i.e., the set of motorbikes and motorcyclist) captured by ten surveillance cameras placed in Brazil's urban traffic scenarios. Furthermore, we evaluate a deep learning model trained using well-known training tricks from the object re-identification literature to present a strong baseline for the motorcycle re-identification (ReID) problem. More importantly, we highlight some crucial problems in this topic as the influence of distractors and the domain shift. Experimental results demonstrate the effectiveness of the strong baseline model with an increase of at least 19.27 p.p. in the rank-1 when compared to the state-of-the-art in the BPReID dataset. Finally, we present some insights regarding the information learned by the strong baseline model when computing the similarities between motorcycle images.

****

Have Fun Storming the Castle(s)!
Connor Anderson, Adam Teuscher, Elizabeth Anderson, Alysia Larsen, Josh Shirley, Ryan Farrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3703-3712

In recent years, large-scale datasets, each typically tailored to a particular problem, have become a critical factor towards fueling rapid progress in the field of computer vision. This paper describes a valuable new dataset that should accelerate research efforts on problems such as fine-grained classification, instance recognition and retrieval, and geolocalization. The dataset is comprised of more than 2400 individual castles, palaces and fortresses from more than 90 countries. The dataset contains more than 770K images in total. This paper details the dataset's construction process, the characteristics including annotations such as location (geotagged latlong and country label), Google Maps link and estimated per-class and per-image difficulty. An experimental section provides baseline experiments for important vision tasks including classification, instance retrieval and geolocalization (estimating the global location from an image's visual appearance).

****

Whose Hand Is This? Person Identification From Egocentric Hand Gestures
Satoshi Tsutsui, Yanwei Fu, David J. Crandall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3399-3408

Recognizing people by faces and other biometrics has been extensively studied in computer vision. But these techniques do not work for identifying the wearer of an egocentric (first-person) camera because that person rarely (if ever) appears in their own first-person view. But while one's own face is not frequently visible, their hands are: in fact, hands are among the most common objects in one's own field of view. It is thus natural to ask whether the appearance and motion patterns of people's hands are distinctive enough to recognize them. In this paper, we systematically study the possibility of Egocentric Hand Identification (EHI) with unconstrained egocentric hand gestures. We explore several different visual cues, including color, shape, skin texture, and depth maps to identify users' hands. Extensive ablation experiments are conducted to analyze the properties of hands that are most distinctive. Finally, we show that EHI can improve generalization of other tasks, such as gesture recognition, by training adversarially to encourage these models to ignore differences between users.

****

Multi-Level Generative Chaotic Recurrent Network for Image Inpainting
Cong Chen, Amos Abbott, Daniel Stilwell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3626-3635

This paper presents a novel multi-level generative chaotic Recurrent Neural Network (RNN) for image inpainting. This technique utilizes a general framework with multiple chaotic RNN that makes learning the image prior from a single corrupted image more robust and efficient. The proposed network utilizes a randomly-initialized process for parameterization, along with a unique quad-directional encoder structure, chaotic state transition, and adaptive importance for multi-level RNN updating. The efficacy of the approach has been validated through multiple experiments. In spite of a much lower computational load, quantitative comparisons reveal that the proposed approach exceeds the performance of several image restoration benchmarks.

****

Self-Supervised Poisson-Gaussian Denoising
Wesley Khademi, Sonia Rao, Clare Minnerath, Guy Hagen, Jonathan Ventura; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2131-2139

We extend the blindspot model for self-supervised denoising to handle Poisson-Gaussian noise and introduce an improved training scheme that avoids hyperparameters and adapts the denoiser to the test data. Self-supervised models for denoising learn to denoise from only noisy data and do not require corresponding clean images, which are difficult or impossible to acquire in some application areas of interest such as low-light microscopy. We introduce a new training strategy to handle Poisson-Gaussian noise which is the standard noise model for microscope images. Our new strategy eliminates hyperparameters from the loss function, which is important in a self-supervised regime where no ground truth data is available to guide hyperparameter tuning. We show how our denoiser can be adapted to the test data to improve performance. Our evaluations on microscope image denoising benchmarks validate our approach.

****

SMPLpix: Neural Avatars From 3D Human Models

Sergey Prokudin, Michael J. Black, Javier Romero; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1810-1819

Recent advances in deep generative models have led to an unprecedented level of realism for synthetically generated images of humans. However, one of the remaining fundamental limitations of these models is the ability to flexibly control the generative process, e.g. change the camera and human pose while retaining the subject identity. At the same time, deformable human body models like SMPL and its successors provide full control over pose and shape but rely on classic computer graphics pipelines for rendering. Such rendering pipelines require explicit mesh rasterization that (a) does not have the potential to fix artifacts or lack of realism in the original 3D geometry and (b) until recently, were not fully incorporated into deep learning frameworks. In this work, we propose to bridge the gap between classic geometry-based rendering and the latest generative networks operating in pixel space. We train a network that directly converts a sparse set of 3D mesh vertices into photorealistic images, alleviating the need for traditional rasterization mechanism. We train our model on a large corpus of human 3D models and corresponding real photos, and show the advantage over conventional differentiable renderers both in terms of the level of photorealism and rendering efficiency.

****

Multi-Modal Trajectory Prediction of NBA Players

Sandro Hauri, Nemanja Djuric, Vladan Radosavljevic, Slobodan Vucetic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1640-1649

National Basketball Association (NBA) players are highly motivated and skilled experts that solve complex decision making problems at every time point during a game. As a step towards understanding how players make their decisions, we focus on their movement trajectories during games. We propose a method that captures the multi-modal behavior of players, where they might consider multiple trajectories and select the most advantageous one. The method is built on an LSTM-based architecture predicting multiple trajectories and their probabilities, trained by a multi-modal loss function that updates the best trajectories. Experiments on large, fine-grained NBA tracking data show that the proposed method outperforms the state-of-the-art. In addition, the results indicate that the approach generates more realistic trajectories and that it can learn individual playing styles of specific players.

****

Multi-Frame Recurrent Adversarial Network for Moving Object Segmentation

Prashant W. Patil, Akshay Dudhane, Subrahmanyam Murala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2302-2311

Moving object segmentation (MOS) in different practical scenarios like weather degraded, dynamic background, etc. videos is a challenging and high demanding task for various computer vision applications. Existing supervised approaches achieve remarkable performance with complicated training or extensive fine-tuning or inappropriate training-testing data distribution. Also, the generalized effect of existing works with completely unseen data is difficult to identify. In this work, the recurrent feature sharing based generative adversarial network is proposed with unseen video analysis. The proposed network comprises of dilated convolution to extract the spatial features at multiple scales. Along with the temporally sampled multiple frames, previous frame output is considered as input to the network. As the motion is very minute between the two consecutive frames, the p revious frame decoder features are shared with encoder features recurrently for current frame foreground segmentation. This recurrent feature sharing of different layers helps the encoder network to learn the hierarchical interactions between the motion and appearance based features. Also, the learning of the proposed network is concentrated in different ways, like disjoint and global training-testing for MOS. An extensive experimental analysis of the proposed network is carried out on two benchmark video datasets with seen and unseen MOS video. Qualitative and quantitative experimental study shows that the proposed network outperforms the existing methods.

****

Autonomous Tracking for Volumetric Video Sequences
Matthew Moynihan, Susana Ruano, Rafael Pages, Aljosa Smolic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1660-1669

As a rapidly growing medium, volumetric video is gaining attention beyond academia, reaching industry and creative communities alike. This brings new challenges to reduce the barrier to entry from a technical and economical point of view. We present a system for robustly and autonomously performing temporally coherent tracking for volumetric sequences, specifically targeting those from sparse setups or with noisy output. Our system will detect and recover missing pertinent geometry across highly incoherent sequences as well as provide users the option of propagating drastic topology edits. In this way, affordable multi-view setups can leverage temporal consistency to reduce processing and compression overheads while also generating more aesthetically pleasing volumetric sequences.

****

EDEN: Multimodal Synthetic Dataset of Enclosed GarDEN Scenes
Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, Theo Gevers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1579-1589

Multimodal large-scale datasets for outdoor scenes are mostly designed for urban driving problems. The scenes are highly structured and semantically different from scenarios seen in nature-centered scenes such as gardens or parks. To promote machine learning methods for nature-oriented applications, such as agriculture and gardening, we propose the multimodal synthetic dataset for Enclosed garDEN scenes (EDEN). The dataset features more than 300K images captured from more than 100 garden models. Each image is annotated with various low/high-level vision modalities, including semantic segmentation, depth, surface normals, intrinsic colors, and optical flow. Experimental results on the state-of-the-art methods for semantic segmentation and monocular depth prediction, two important tasks in computer vision, show positive impact of pre-training deep networks on our dataset for unstructured natural scenes. The dataset and related materials will be available at https://lhoangan.github.io/eden.

****

Single Image Reflection Removal With Edge Guidance, Reflection Classifier, and Recurrent Decomposition
Ya-Chu Chang, Chia-Ni Lu, Chia-Chi Cheng, Wei-Chen Chiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2033-2042

Removing undesired reflection from an image captured through a glass window is a notable task in computer vision. In this paper, we propose a novel model with auxiliary techniques to tackle the problem of single image reflection removal. Our model takes a reflection contaminated image as input, and decomposes it into the reflection layer and the transmission layer. In order to ensure quality of the transmission layer, we introduce three auxiliary techniques into our architecture, including the edge guidance, a reflection classifier, and the recurrent decomposition. The contributions and the efficacy of these techniques are investigated and verified in the ablation study. Furthermore, in comparison to the state-of-the-art baselines of reflection removal, both quantitative and qualitative results demonstrate that our proposed method is able to deal with different kinds of images, achieving the best results in average.

****

Utilizing Every Image Object for Semi-Supervised Phrase Grounding

Haidong Zhu, Arka Sadhu, Zhaoheng Zheng, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2210-2219

Phrase grounding models localize an object in the image given a referring expression. The annotated language queries available during training are limited, which also limits the variations of language combinations that a model can see during training. In this paper, we study the case applying objects without labeled queries for training the semi-supervised phrase grounding. We propose to use learned location and subject embedding predictors (LSEP) to generate the corresponding language embeddings for objects lacking annotated queries in the training set. With the assistance of the detector, we also apply LSEP to train a grounding model on images without any annotation. We evaluate our method based on MAttNet on three public datasets: RefCOCO, RefCOCO+, and RefCOCOg. We show that our predictors allow the grounding system to learn from the objects without labeled queries and improve accuracy by 34.9% relatively with the detection results.

****

Guided Attentive Feature Fusion for Multispectral Pedestrian Detection

Heng Zhang, Elisa Fromont, Sebastien Lefevre, Bruno Avignon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 72-80

Multispectral image pairs can provide complementary visual information, making pedestrian detection systems more robust and reliable. To benefit from both RGB and thermal IR modalities, we introduce a novel attentive multispectral feature fusion approach. Under the guidance of the inter- and intra-modality attention modules, our deep learning architecture learns to dynamically weigh and fuse the multispectral features. Experiments on two public multispectral object detection datasets demonstrate that the proposed approach significantly improves the detection accuracy at a low computation cost.

****

Learning of Low-Level Feature Keypoints for Accurate and Robust Detection

Suwichaya Suwanwimolkul, Satoshi Komorita, Kazuyuki Tasaka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2262-2271

Joint learning of feature descriptor and detector has offered promising 3D reconstruction results; however, they often lack the low-level feature awareness, which causes low accuracy in matched keypoint locations. The others employed fixed operations to select the keypoints, but the selected keypoints may not correspond to the descriptor matching. To address these problems, we propose the supervised learning of keypoint detection with low-level features. Our detector is a single CNN layer extended from the descriptor backbone, which can be jointly learned with the descriptor for maximizing the descriptor matching. This results in a state-of-the-art 3D reconstruction, especially on improving reprojection error, and the highest accuracy in keypoint detection and matching on benchmark datasets. We also present a dedicated study on evaluation metrics to measure the accuracy of keypoint detection and matching.

****

Receptive Field Size Optimization With Continuous Time Pooling

Dora Babicz, Soma Kontar, Mark Peto, Andras Fulop, Gergely Szabo, Andras Horvath; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1449-1458

Pooling operation is a cornerstone element of convolutional neural networks. These elements generate receptive fields for neurons, in which local perturbations should have minimal effect on the output activations, increasing robustness and invariance of the whole network. In this paper we will present an altered version of the most commonly applied method, maximum pooling, where pooling in theory is substituted by a continuous time differential equation, which generates a location sensitive pooling operation, which is more similar to biological receptive fields. We will present how this continuous method can be approximated numerica lly using discreet operations which fit ideally on a GPU. In our approach the hyperparameter kernel size is substituted by diffusion strength which is a continuous value, this way it can be optimized by gradient descent algorithms. We will evaluate the effect of continuous pooling on accuracy using commonly applied network architectures and datasets.

****

AI on the Bog: Monitoring and Evaluating Cranberry Crop Risk

Peri Akiva, Benjamin Planche, Aditi Roy, Kristin Dana, Peter Oudemans, Michael Mars; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2493-2502

Machine vision for precision agriculture has attracted considerable research interest in recent years. The goal of this paper is to develop an end-end cranberry health monitoring system to enable and support real time cranberry over-heating assessment to facilitate informed decisions that may sustain the economic viability of the farm. Toward this goal, we propose two main deep learning-based modules for: 1) cranberry fruit segmentation to delineate the exact fruit regions in the cranberry field image that are exposed to sun, 2) prediction of cloud coverage conditions to estimate the inner temperature of exposed cranberries We develop drone-based field data and ground-based sky data collection systems to collect video imagery at multiple time points for use in crop health analysis. Extensive evaluation on the data set shows that it is possible to predict exposed fruit's inner temperature with high accuracy (0.02% MAPE) when irradiance is predicted with 5.59-19.84% MAPE in the 5-20 minutes time horizon. With 62.54% mIoU for segmentation and 13.46 MAE for counting accuracies in exposed fruit identification, this system is capable of giving informed feedback to growers to take precautionary action (e.g., irrigation) in identified crop field regions with higher risk of sunburn in the near future. Though this novel system is applied for cranberry health monitoring, it represents a pioneering step forward in efficiency for farming and is useful in precision agriculture beyond the problem of cranberry overheating.

****

Exploration of Spatial and Temporal Modeling Alternatives for HOI

Rishabh Dabral, Srijon Sarkar, Sai Praneeth Reddy, Ganesh Ramakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2281-2290

Human-Object Interaction detection from a video clip can be considered as a special case of video-based Visual-Relationship Detection wherein the subject must be a human. Specifically, it involves detecting the humans and objects in the clip as well as the interactions between them. Conventionally, the problem has been formulated as a space-time graph inference problem over the video clip features. In this work, we explore alternate spatial approaches for detecting Human-Object Interactions. We consider a hierarchical setup that decouples spatial and temporal aspects of the problem and analyse the impacts of a variety of design choices for the spatial networks. Particularly, to capture spatial relationships in the scene, we analyze the effectiveness of the traditionally used Graph Convolutional Networks against Convolutional Networks and Capsule Networks. Unlike current approaches, we avoid using ground truth data like depth maps or 3D human pose during inference, thus increasing generalization across non-RGBD datasets as well. We demonstrate a comprehensive analysis of the exploration, both quantitatively and qualitatively, while achieving state-of-the-art results in human-object interaction detection (88.9% and 92.6%) and anticipation tasks of CAD-120 and competitive results on image based HOI detection (47.2%) in V-COCO dataset, setting a new benchmark for visual features based approaches.

****

Mask Selection and Propagation for Unsupervised Video Object Segmentation

Shubhika Garg, Vidit Goel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1680-1690

In this work we present a novel approach for Unsupervised Video Object Segmentation, that is automatically generating instance level segmentation masks for salient objects and tracking them in a video. We efficiently handle problems present in existing methods such as drift while temporal propagation, tracking and addition of new objects. To this end, we propose a novel idea of improving masks in an online manner using ensemble of criteria whose task is to inspect the quality of masks. We introduce a novel idea of assessing mask quality using a neural network called Selector Net. The proposed network is trained is such way that it is generalizes across various datasets. Our proposed method is able to limit the noise accumulated along the video, giving state of the art result on Davis 2019 Unsupervised challenge dataset with J&F mean 61.6%. We also tested on datasets such as FBMS and SegTrack V2 and performed better or on par compared to the other methods.

****

Where to Look?: Mining Complementary Image Regions for Weakly Supervised Object Localization

Sadbhavana Babar, Sukhendu Das; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1010-1019

Humans possess an innate capability of recognizing objects and their corresponding parts and confine their attention to that location in a visual scene where the object is spatially present. Recently, efforts to train machines to mimic this ability of humans in the form of weakly supervised object localization, using training labels only at the image-level, have garnered a lot of attention. Nonetheless, one of the well-known problems that most of the existing methods suffer from is localizing only the most discriminative part of an object. Such methods provide very little or no focus on other pertinent parts of the object. In this paper, we propose a novel way of scrupulously localizing objects using training with labels as for the entire image by mining information from complementary regions in an image. Primarily, we adapt to regional dropout at complementary spatial locations to create two intermediate images. With the help of a novel Channel-wise Assisted Attention Module (CAAM) coupled with a Spatial Self-Attention Module (SSAM), we parallely train our model to leverage the information from complementary image regions for excellent localization. Finally, we fuse the attention maps generated by the two classifiers using our Attention-based Fusion Loss. Several experimental studies manifest the superior performance of our proposed approach. Our method demonstrates a significant increase in localization performance over the existing state-of-the-art methods on CUB-200-2011 and ILSVRC 2016 datasets.

****

RefineLoc: Iterative Refinement for Weakly-Supervised Action Localization

Alejandro Pardo, Humam Alwassel, Fabian Caba, Ali Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3319-3328

Video action detectors are usually trained using datasets with fully-supervised temporal annotations. Building such datasets is an expensive task. To alleviate this problem, recent methods have tried to leverage weak labeling, where videos are untrimmed and only a video-level label is available. In this paper, we propose RefineLoc, a novel weakly-supervised temporal action localization method. RefineLoc uses an iterative refinement approach by estimating and training on snippet-level pseudo ground truth at every iteration. We show the benefit of this iterative approach and present an extensive analysis of five different pseudo ground truth generators. We show the effectiveness of our model on two standard action datasets, ActivityNet v1.2 and THUMOS14. RefineLoc shows competitive results with the state-of-the-art in weakly-supervised temporal localization. Additionally, our iterative refinement process is able to significantly improve the performance of two state-of-the-art methods, setting a new state-of-the-art on THUMOS14.

****

MinkLoc3D: Point Cloud Based Large-Scale Place Recognition

Jacek Komorowski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1790-1799

The paper presents a learning-based method for computing a discriminative 3D point cloud descriptor for place recognition purposes. Existing methods, such as Po intNetVLAD, are based on unordered point cloud representation. They use PointNet as the first processing step to extract local features, which are later aggregated into a global descriptor. The PointNet architecture is not well suited to capture local geometric structures. Thus, state-of-the-art methods enhance vanilla PointNet architecture by adding different mechanism to capture local contextual information, such as graph convolutional networks or using hand-crafted features. We present an alternative approach, dubbed MinkLoc3D, to compute a discriminative 3D point cloud descriptor, based on a sparse voxelized point cloud representation and sparse 3D convolutions. The proposed method has a simple and efficient architecture. Evaluation on standard benchmarks proves that MinkLoc3D outperforms current state-of-the-art. Our code is publicly available on the project website.

****

Towards Precise Intra-Camera Supervised Person Re-Identification

Menglin Wang, Baisheng Lai, Haokun Chen, Jianqiang Huang, Xiaojin Gong, Xian-Sheng Hua; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3229-3238

Intra-camera supervision (ICS) for person re-identification (Re-ID) assumes that identity labels are independently annotated within each camera view and no inter-camera identity association is labeled. It is a new setting proposed recently to reduce the burden of annotation while expect to maintain desirable Re-ID performance. However, the lack of inter-camera labels makes the ICS Re-ID problem much more challenging than the fully supervised counterpart. By investigating the characteristics of ICS, this paper proposes jointly learned camera-specific non-parametric classifiers, together with a hybrid mining quintuplet loss, to perform intra-camera learning. Then, an inter-camera learning module consisting of a graph-based ID association step and a Re-ID model updating step is conducted. Extensive experiments on three large-scale Re-ID datasets show that our approach outperforms all existing ICS works by a great margin. Our approach performs even comparable to state-of-the-art fully supervised methods in two of the datasets.

****

Structured Visual Search via Composition-Aware Learning

Mert Kilickaya, Arnold W.M. Smeulders; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1701-1710

This paper studies visual search using structured queries. The structure is in the form of a 2D composition that encodes the position and the category of the objects. The transformation of the position and the category of the objects leads to a continuous-valued relationship between visual compositions, which carries highly beneficial information, although not leveraged by previous techniques. To that end, in this work, our goal is to leverage these continuous relationships by using the notion of symmetry in equivariance. Our model output is trained to change symmetrically with respect to the input transformations, leading to a sensitive feature space. Doing so leads to a highly efficient search technique, as our approach learns from fewer data using a smaller feature space. Experiments on two large-scale benchmarks of MS-COCO and HICO-DET demonstrates that our approach leads to a considerable gain in the performance against competing techniques.

****

Learning to Distill Convolutional Features Into Compact Local Descriptors

Jongmin Lee, Yoonwoo Jeong, Seungwook Kim, Juhong Min, Minsu Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 898-908

Extracting local descriptors or features is an essential step in solving image matching problems. Recent methods in the literature mainly focus on extracting effective descriptors, without much attention to the size of the descriptors. In this work, we study how to learn a compact yet effective local descriptor. The proposed method distills multiple intermediate features of a pretrained convolutional neural network to encode different levels of visual information from local textures to non-local semantics, resulting in local descriptors with a designated dimension. Experiments on standard benchmarks for semantic correspondence show that it achieves significantly improved performance over existing models, with u

p to a 100 times smaller size of descriptors. Furthermore, while trained on a small-sized dataset for semantic correspondence, the proposed method also generalizes well to other image matching tasks, performing comparable to the state of the art on wide-baseline matching and visual localization benchmarks.
****

Spatial Context-Aware Self-Attention Model for Multi-Organ Segmentation

Hao Tang, Xingwei Liu, Kun Han, Xiaohui Xie, Xuming Chen, Huang Qian, Yong Liu, Shanlin Sun, Narisu Bai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 939-949

Multi-organ segmentation is one of most successful applications of deep learning in medical image analysis. Deep convolutional neural nets (CNNs) have shown great promise in achieving clinically applicable image segmentation performance on CT or MRI images. State-of-the-art CNN segmentation models apply either 2D or 3D convolutions on input images, with pros and cons associated with each method: 2D convolution is fast, less memory-intensive but inadequate for extracting3D contextual information from volumetric images, while the opposite is true for 3D convolution. To fit a 3D CNN model on CT or MRI images on commodity GPUs, one usually has to either downsample input images or use cropped local regions as inputs, which limits the utility of3D models for multi-organ segmentation. In this work, we propose a new framework for combining 3D and 2D models, in which the segmentation is realized through high-resolution 2D convolutions, but guided by spatial contextual information extracted from a low-resolution 3D model.We implement a self-attention mechanism to control which 3D features should be used to guide 2D segmentation. Our model is light on memory usage but fully equipped to take 3D contextual information into account.Experiments on multiple organ segmentation datasets demonstrate that by taking advantage of both 2D and 3D models, our method consistently outperforms existing 2D and 3D models in organ segmentation accuracy, while being able to directly take raw whole-volume image data as inputs.
****

2D to 3D Medical Image Colorization

Aradhya Neeraj Mathur, Apoorv Khattar, Ojaswa Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2847-2856

Colorization involves the synthesis of colors while preserving structural content as well as the semantics of the target image. This is a well-explored problem in 2D with many state-of-the-art solutions. We explore a new challenge in the field of colorization where we aim at colorizing multi-modal 3D medical data using style exemplars. To the best of our knowledge, this work is the first of its kind and so we discuss the full pipeline in detail and the challenges that it brings for 3D medical data. The colorization of medical MRI volume also entails modality conversion that highlights the robustness of our approach in handling multi-modal data.
****

IGSSTRCF: Importance Guided Sparse Spatio-Temporal Regularized Correlation Filters for Tracking

Monika Jain, A. V. Subramanyam, Simon Denman, Sridha Sridharan, Clinton Fookes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2775-2784

This paper proposes a novel Importance Guided Sparse Spatio-Temporal Regularization based Correlation Filter (IGSSTRCF) tracker. Our formulation explicitly models the variations in the correlation filters and associated spatial weights in successive frames. By imposing a sparsity penalty on these variations, the formulation ensures that only relevant changes are incorporated during updates. This results in more robust filter coefficients that minimize the tracking drift. The IGSSTRCF also includes an adaptive channel importance estimation strategy that assigns an importance weight to each feature channel during training. The proposed formulation is efficiently solved via the alternating direction method of multipliers. A comparative analysis is shown on TC128, UAV123, VOT-2017, and VOT-2019 datasets; and we present an ablation study to demonstrate the contribution of each component of the IGSSTRCF. It is observed that we outperform several state-of-the-art trackers and each component of the proposed IGSSTRCF contributes positively towards tracker performance.

****

SLAM in the Field: An Evaluation of Monocular Mapping and Localization on Challenging Dynamic Agricultural Environment

Fangwen Shu, Paul Lesur, Yaxu Xie, Alain Pagani, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1761-1771

This paper demonstrates a system capable of combining a sparse, indirect, monocular visual SLAM, with both offline and real-time Multi-View Stereo (MVS) reconstruction algorithms. This combination overcomes many obstacles encountered by autonomous vehicles or robots employed in agricultural environments, such as overly repetitive patterns, need for very detailed reconstructions, and abrupt movements caused by uneven roads. Furthermore, the use of a monocular SLAM makes our system much easier to integrate with an existing device, as we do not rely on a LiDAR (which is expensive and power consuming), or stereo camera (whose calibration is sensitive to external perturbation e.g. camera being displaced). To the best of our knowledge, this paper presents the first evaluation results for monocular SLAM, and our work further explores unsupervised depth estimation on this specific application scenario by simulating RGB-D SLAM to tackle the scale ambiguity, and shows our approach produces reconstructions that are helpful to various agricultural tasks. Moreover, we highlight that our experiments provide meaningful insight to improve monocular SLAM systems under agricultural settings.

****

Self-Supervised Visual-LiDAR Odometry With Flip Consistency

Bin Li, Mu Hu, Shuling Wang, Lianghao Wang, Xiaojin Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3844-3852

Most learning-based methods estimate ego-motion by utilizing visual sensors, which suffer from dramatic lighting variations and textureless scenarios. In this paper, we incorporate sparse but accurate depth measurements obtained from lidars to overcome the limitation of visual methods. To this end, we design a self-supervised visual-lidar odometry (Self-VLO) framework. It takes both monocular images and sparse depth maps projected from 3D lidar points as input, and produces pose and depth estimations in an end-to-end learning manner, without using any ground truth labels. To effectively fuse two modalities, we design a two-pathway encoder to extract features from visual and depth images and fuse the encoded features with those in decoders at multiple scales by our fusion module. We also adopt a siamese architecture and design an adaptively weighted flip consistency loss to facilitate the selfsupervised learning of our VLO. Experiments on the KITTI odometry benchmark show that the proposed approach outperforms all self-supervised visual or lidar odometries. It also performs better than fully supervised VOs, demonstrating the power of fusion.

****

DB-GAN: Boosting Object Recognition Under Strong Lighting Conditions

Luca Minciullo, Fabian Manhardt, Kei Yoshikawa, Sven Meier, Federico Tombari, Norimasa Kobori; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2939-2949

Driven by deep learning, object recognition has recently made a tremendous leap forward. Nonetheless, its accuracy often still suffers from several sources of variation that can be found in real-world images. Some of the most challenging variation are induced by changing lighting conditions. This paper presents a novel approach for tackling bright-ness variation in the domain of 2D object detection and 6D object pose estimation. Existing works aiming at improving robustness towards different lighting conditions are of-ten grounded on classical computer vision contrast normalisation techniques or the acquisition of large amounts of an-notated data in order to achieve invariance during training.While the former cannot generalise well to a wide range of illumination conditions, the latter is neither practical nor scalable. Hence, we propose the usage of Generative Adversarial Network in order to learn how to normalise the illumination of an input im age. Thereby, the generator is explicitly designed to normalise illumination in images soto enhance the object recognition performance. Extensive evaluations demonstrate that leveraging the generated data can significantly enhance the detection performance, out-performing all other state-of-the-art methods. We further constitute a natural extension focusing on white balance variations and introduce a new dataset for evaluation.

****

Deep Preset: Blending and Retouching Photos With Color Style Transfer

Man M. Ho, Jinjia Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2113-2121

End-users, without knowledge in photography, desire to beautify their photos to have a similar color style as a well-retouched reference. However, recent works in image style transfer are overused. They usually synthesize undesirable results due to transferring exact colors to the wrong destination. It becomes even worse in sensitive cases such as portraits. In this work, we concentrate on learning low-level image transformation, especially color-shifting methods, rather than contextual features matching, then present a novel supervised approach for color style transfer. Furthermore, we propose a color style transfer named Deep Preset designed to 1) generalize the features representing the color transformation from content with natural colors to retouched reference, then blend it into the contextual features of content, 2) predict hyper-parameters (settings or preset) of the applied low-level color transformation methods, 3) stylize content image to have a similar color style as reference. We script Lightroom, a powerful tool in editing photos, to generate 600,000 training samples using 1,200 images from the Flick2K dataset and 500 user-generated presets with 69 settings. Experimental results show that our Deep Preset outperforms the previous works in color style transfer quantitatively and qualitatively. Our work is available at https://minhmanho.github.io/deep_preset/.

****

Benefiting From Bicubically Down-Sampled Images for Learning Real-World Image Super-Resolution

Mohammad Saeed Rad, Thomas Yu, Claudiu Musat, Hazim Kemal Ekenel, Behzad Bozorgtabar, Jean-Philippe Thiran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1590-1599

Super-resolution (SR) has traditionally been based on pairs of high-resolution images (HR) and their low-resolution (LR) counterparts obtained artificially with bicubic downsampling. However, in real-world SR, there is a large variety of realistic image degradations and analytically modeling these realistic degradations can prove quite difficult. In this work, we propose to handle real-world SR by splitting this ill-posed problem into two comparatively more well-posed steps. First, we train a network to transform real LR images to the space of bicubically downsampled images in a supervised manner, by using both real LR/HR pairs and synthetic pairs. Second, we take a generic SR network trained on bicubically downsampled images to super-resolve the transformed LR image. The first step of the pipeline addresses the problem by registering the large variety of degraded images to a common, well understood space of images. The second step then leverages the already impressive performance of SR on bicubically downsampled images, sidestepping the issues of end-to-end training on datasets with many different image degradations. We demonstrate the effectiveness of our proposed method by comparing it to recent methods in real-world SR and show that our proposed approach outperforms the state-of-the-art works in terms of both qualitative and quantitative results, as well as results of an extensive user study conducted on several real image datasets.

****

CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, Heung-Kyu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 375-384

Detecting and localizing image splicing has become essential to fight against ma licious forgery. A major challenge to localize spliced areas is to discriminate between authentic and tampered regions with intrinsic properties such as compression artifacts. We propose CAT-Net, an end-to-end fully convolutional neural network including RGB and DCT streams, to learn forensic features of compression artifacts on RGB and DCT domains jointly. Each stream considers multiple resolutions to deal with spliced object's various shapes and sizes. The DCT stream is pretrained on double JPEG detection to utilize JPEG artifacts. The proposed method outperforms state-of-the-art neural networks for localizing spliced regions in JPEG or non-JPEG images.

****

Fusion Learning Using Semantics and Graph Convolutional Network for Visual Food Recognition

Heng Zhao, Kim-Hui Yap, Alex Chichung Kot; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1711-1720

Food-related applications and services are essential for the health and well-being of people. With the rapid development of social networks and mobile devices, food images captured by people can offer rich knowledge about the food and also necessary dietary assistance for people that require special care. Known food recognition frameworks and approaches in computer vision have heavy reliance on many-shot training of a deep network on existing large-scale food datasets. However, it is common for many food categories that it is difficult to collect enough images for training. Traditional few-shot learning is unable to properly address the problem due to the complex characteristics and large variations of food images, and most few-shot frameworks cannot perform classification for many-shot and few-shot categories at the same time. In this paper, we propose a new fusion learning framework for food recognition. It unifies many-shot and few-shot under a single framework, by leveraging on extracted image representations and context sensitive semantic embeddings. Further, considering food categories are often correlated to each other for many commonalities such as same ingredients, cooking methods, the fusion learning framework utilizes a Graph Convolutional Network (GCN) to capture the inter-class relations between both image representations and semantic embeddings of different food categories. The final output fusion classifier will be more robust and discriminative. Comprehensive experimental results on two popular food benchmarks have shown the proposed framework achieves the state-of-the-art fusion performance.

****

Exploiting Spatial Relation for Reducing Distortion in Style Transfer

Jia-Ren Chang, Yong-Sheng Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1209-1217

The power of convolutional neural networks in arbitrary style transfer has been amply demonstrated; however, existing stylization methods tend to generate spatially inconsistent results with noticeable artifacts. One solution to this problem involves the application of a segmentation mask or affinity-based image matting to preserve spatial information related to image content. The main idea of this work is to model spatial relation between content image pixels and thus to maintain this relationship in stylization for reducing artifacts. The proposed network architecture is called spatial relation-augmented VGG (SRVGG), in which long-range spatial dependency is modeled by a spatial relation module. Based on this spatial information extracted from SRVGG, we design a novel relation loss which can minimize the difference of spatial dependency between content images and stylizations. We evaluate the proposed framework on both optimization-based and feedforward-based style transfer methods. The effectiveness of SRVGG in stylization is demonstrated by generating stylized images of high quality and spatial consistency without the need for segmentation masks or affinity-based image matting. The quantitative evaluation also suggests that the proposed framework achieve better performance compared with other methods.

****

Overcomplete Deep Subspace Clustering Networks

Jeya Maria Jose Valanarasu, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 746-755

Deep Subspace Clustering Networks (DSC) provide an efficient solution to the problem of unsupervised subspace clustering by using an undercomplete deep auto-encoder with a fully-connected layer to exploit the self expressiveness property. This method uses undercomplete representations of the input data which makes it not so robust and more dependent on pre-training. To overcome this, we propose a simple yet efficient alternative method - Overcomplete Deep Subspace Clustering Networks (ODSC) where we use overcomplete representations for subspace clustering. In our proposed method, we fuse the features from both undercomplete and overcomplete auto-encoder networks before passing them through the self-expressive layer thus enabling us to extract a more meaningful and robust representation of the input data for clustering. Experimental results on four benchmark datasets show the effectiveness of the proposed method over DSC and other clustering methods in terms of clustering error. Our method is also not as dependent as DSC is on where pre-training should be stopped to get the best performance and is also more robust to noise.

****

Asymmetric Contextual Modulation for Infrared Small Target Detection

Yimian Dai, Yiquan Wu, Fei Zhou, Kobus Barnard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 950-959

Single-frame infrared small target detection remains a challenge not only due to the scarcity of intrinsic target characteristics but also because of lacking a public dataset. In this paper, we first contribute an open dataset with high-quality annotations to advance the research in this field. We also propose an asymmetric contextual modulation module specially designed for detecting infrared small targets. To better highlight small targets, besides a top-down global contextual feedback, we supplement a bottom-up modulation pathway based on point-wise channel attention for exchanging high-level semantics and subtle low-level details. We report ablation studies and comparisons to state-of-the-art methods, where we find that our approach performs significantly better. Our dataset and code are available online.

****

Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, Nancy Xin Ru Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 697-706

Documents are often the format of choice for knowledge sharing and preservation in business and science, within which are tables that capture most of the critical data. Unfortunately, most documents are stored and distributed as PDF or scanned images, which fail to preserve table formatting. Recent vision-based deep learning approaches have been proposed to address this gap, but most still cannot achieve state-of-the-art results. We present Global Table Extractor (GTE), a vision-guided systematic framework for joint table detection and cell structured recognition, which could be built on top of any object detection model. With GTE-Table, we invent a new penalty based on the natural cell containment constraint of tables to train our table network aided by cell location predictions. GTE-Cell is a new hierarchical cell detection network that leverages table styles. Further, we design a method to automatically label table and cell structure in existing documents to cheaply create a large corpus of training and test data. We use this to enhance PubTabNet with cell labels and create FinTabNet, real-world and complex scientific and financial datasets with detailed table structure annotations to help train and test structure recognition. Our deep learning framework surpasses previous state-of-the-art results on the ICDAR 2013 and ICDAR 2019 table competition test dataset in both table detection and cell structure recognition. Further experiments demonstrate a greater than 45% improvement in cell structure recognition when compared to a vanilla RetinaNet object detection model in our new out-of-domain financial dataset (Fintabnet).

****

Active Learning for Bayesian 3D Hand Pose Estimation

Razvan Caramalau, Binod Bhattarai, Tae-Kyun Kim; Proceedings of the IEEE/CVF Win ter Conference on Applications of Computer Vision (WACV), 2021, pp. 3419-3428

We propose a Bayesian approximation to a deep learning architecture for 3D hand pose estimation. Through this framework, we explore and analyse the two types of uncertainties that are influenced either by data or by the learning capability. Furthermore, we draw comparisons against the standard estimator over three popular benchmarks. The first contribution lies in outperforming the baseline while in the second part we address the active learning application. We also show that with a newly proposed acquisition function, our Bayesian 3D hand pose estimator obtains lowest errors with the least amount of data. The underlying code is publicly available at https://github.com/razvancaramalau/al_bhpe.

****

Red Carpet to Fight Club: Partially-Supervised Domain Transfer for Face Recognition in Violent Videos

Yunus Can Bilge, Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, Pinar Duygulu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3358-3369

In many real-world problems, there is typically a large discrepancy between the characteristics of data used in training versus deployment. A prime example is the analysis of aggression videos: in a criminal incidence, typically suspects need to be identified based on their clean portrait-like photos, instead of their prior video recordings. This results in three major challenges; large domain discrepancy between violence videos and ID-photos, the lack of video examples for most individuals and limited training data availability. To mimic such scenarios, we formulate a realistic domain-transfer problem, where the goal is to transfer the recognition model trained on clean posed images to the target domain of violent videos, where training videos are available only for a subset of subjects. To this end, we introduce the "WildestFaces" dataset, tailored to study cross-domain recognition under a variety of adverse conditions. We divide the task of transferring a recognition model from the domain of clean images to the violent videos into two sub-problems and tackle them using (i) stacked affine-transforms for classifier-transfer, (ii) attention-driven pooling for temporal-adaptation. We additionally formulate a self attention based model for domain-transfer. We establish a rigorous evaluation protocol for this "clean-to-violent" recognition task, and present a detailed analysis of the proposed dataset and the methods. Our experiments highlight the unique challenges introduced by the Wildest-Faces dataset and the advantages of the proposed approach.

****

TracKlinic: Diagnosis of Challenge Factors in Visual Tracking

Heng Fan, Fan Yang, Peng Chu, Yuewei Lin, Lin Yuan, Haibin Ling; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 970-979

Generic visual object tracking is difficult due to many challenge factors (e.g., occlusion, blur, etc.). Each of these factors may cause serious problems for a tracker, and when they work together can make things even more complicated. Despite a great amount of efforts devoted to understanding the behavior of trackers, reliable and quantifiable ways for studying the per factor tracking behavior remain barely available. Addressing this issue, in this paper we contribute to the community a tracking diagnosis toolkit, TracKlinic, for diagnosis of challenge factors of tracking algorithms. TracKlinic consists of two novel components focusing on the data and analysis aspects, respectively. For the data component, we carefully prepare a set of 2,390 annotated videos, each involving one and only one major challenge factor. When analyzing an algorithm for a specific challenge factor, such one-factor-per-sequence rule greatly inhibits the disturbance from other factors and consequently leads to more faithful analysis. For the analysis component, given the tracking results on all sequences, it investigates the behavior of the tracker under each individual factor and generates the report automatically. With TracKlinic, a thorough study is conducted on ten state-of-the-art trackers on nine challenge factors (including two compound ones). The results suggest that, heavy shape variation and occlusion are the two most challenging factors faced by most trackers. Besides, out-of-view, though does not happen frequ ently, is often fatal. By sharing TracKlinic, we expect to make it much easier for diagnosing tracking algorithms, and to thus facilitate developing better ones.

****

SynDistNet: Self-Supervised Monocular Fisheye Camera Distance Estimation Synergized With Semantic Segmentation for Autonomous Driving

Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, Patrick Mader; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 61-71

State-of-the-art self-supervised learning approaches for monocular depth estimation usually suffer from scale ambiguity. They do not generalize well when applied on distance estimation for complex projection models such as in fisheye and omnidirectional cameras. This paper introduces a novel multi-task learning strategy to improve self-supervised monocular distance estimation on fisheye and pinhole camera images. Our contribution to this work is threefold: Firstly, we introduce a novel distance estimation network architecture using a self-attention based encoder coupled with robust semantic feature guidance to the decoder that can be trained in a one-stage fashion. Secondly, we integrate a generalized robust loss function, which improves performance significantly while removing the need for hyperparameter tuning with the reprojection loss. Finally, we reduce the artifacts caused by dynamic objects violating static world assumptions using a semantic masking strategy. We significantly improve upon the RMSE of previous work on fisheye by a 25% reduction in RMSE. As there is little work on fisheye cameras, we evaluated the proposed method on KITTI using a pinhole model. We achieved state-of-the-art performance among self-supervised methods without requiring an external scale estimation.

****

Video Captioning of Future Frames

Mehrdad Hosseinzadeh, Yang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 980-989

Being able to anticipate and describe what may happen in the future is a fundamental ability for humans. Given a short clip of a scene about "a person is sitting behind a piano", humans can describe what will happen afterward, i.e. "the person is playing the piano". In this paper, we consider the task of captioning future events to assess the performance of intelligent models on anticipation and video description generation tasks simultaneously. More specifically, given only the frames relating to an occurring event (activity), the goal is to generate a sentence describing the most likely next event in the video. We tackle the problem by first predicting the next event in the semantic space of convolutional features, then fusing contextual information into those features, and feeding them to a captioning module. Departing from using recurrent units allows us to train the network in parallel. We compare the proposed method with a baseline and an oracle method on the ActivityNetCaptions dataset. Experimental results demonstrate that the proposed method outperforms the baseline and is comparable to the oracle method. We perform additional ablation study to further analyze our approach.

****

CIT-GAN: Cyclic Image Translation Generative Adversarial Network With Application in Iris Presentation Attack Detection

Shivangi Yadav, Arun Ross; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2412-2421

In this work, we propose a novel Cyclic Image Translation Generative Adversarial Network (CIT-GAN) for multi-domain style transfer. To facilitate this, we introduce a Styling Network that has the capability to learn style characteristics of each domain represented in the training dataset. The Styling Network helps the generator to drive the translation of images from a source domain to a reference domain and generate synthetic images with style characteristics of the reference domain. The learned style characteristics for each domain depend on both the style loss and domain classification loss. This induces variability in style characteristics within each domain. The proposed CIT-GAN is used in the context of i ris presentation attack detection (PAD) to generate synthetic presentation attack (PA) samples for classes that are under-represented in the training set. Evaluation using current state-of-the-art iris PAD methods demonstrates the efficacy of using such synthetically generated PA samples for training PAD methods. Further, the quality of the synthetically generated samples is evaluated using Frechet Inception Distance (FID) score. Results show that the quality of synthetic images generated by the proposed method is superior to that of other competing methods, including StarGan.

****

Representation Learning Through Latent Canonicalizations

Or Litany, Ari Morcos, Srinath Sridhar, Leonidas Guibas, Judy Hoffman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 645-654

We seek to learn a representation on a large annotated data source that generalizes to a target domain using limited new supervision. Many prior approaches to this problem have focused on learning "disentangled" representations so that as individual factors vary in a new domain, only a portion of the representation need be updated. In this work, we seek the generalization power of disentangled representations, but relax the requirement of explicit latent disentanglement and instead encourage linearity of individual factors of variation by requiring them to be manipulable by learned linear transformations. We dub these transformations latent canonicalizers, as they aim to modify the value of a factor to a pre-determined (but arbitrary) canonical value (e.g., recoloring the image foreground to black). Assuming a source domain with access to meta labels specifying the factors of variation within an image, we demonstrate experimentally that our method helps reduce the number of observations needed to generalize to a similar target domain when compared to a number of supervised baselines.

****

A Weakly Supervised Consistency-Based Learning Method for COVID-19 Segmentation in CT Images

Issam Laradji, Pau Rodriguez, Oscar Manas, Keegan Lensink, Marco Law, Lironne Kurzman, William Parker, David Vazquez, Derek Nowrouzezahrai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2453-2462

Coronavirus Disease 2019 (COVID-19) has spread aggressively across the world causing an existential health crisis. Thus, having a system that automatically detects COVID-19 in tomography (CT) images can assist in quantifying the severity of the illness. Unfortunately, labelling chest CT scans requires significant domain expertise, time, and effort. We address these labelling challenges by only requiring point annotations, a single pixel for each infected region on a CT image. This labeling scheme allows annotators to label a pixel in a likely infected region, only taking 1-3 seconds, as opposed to 10-15 seconds to segment a region. Conventionally, segmentation models train on point-level annotations using the cross-entropy loss function on these labels. However, these models often suffer from low precision. Thus, we propose a consistency-based (CB) loss function that encourages the output predictions to be consistent with spatial transformations of the input images. The experiments on 3 open-source COVID-19 datasets show that this loss function yields significant improvement over conventional point-level loss functions and almost matches the performance of models trained with full supervision with much less human effort. Code is available at: https://github.com/IssamLaradji/covid19_weak_supervision.

****

Generalized Object Detection on Fisheye Cameras for Autonomous Driving: Dataset, Representations and Baseline

Hazem Rashed, Eslam Mohamed, Ganesh Sistu, Varun Ravi Kumar, Ciaran Eising, Ahmad El-Sallab, Senthil Yogamani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2272-2280

Object detection is a comprehensively studied problem in autonomous driving. However, it has been relatively less explored in the case of fisheye cameras. The standard bounding box fails in fisheye cameras due to the strong radial distortio n, particularly in the image's periphery. We explore better representations like oriented bounding box, ellipse, and generic polygon for object detection in fisheye images in this work. We use the IoU metric to compare these representations using accurate instance segmentation ground truth. We design a novel curved bounding box model that has optimal properties for fisheye distortion models. We also design a curvature adaptive perimeter sampling method for obtaining polygon vertices, improving relative mAP score by 4.9 % compared to uniform sampling. Overall, the proposed polygon model improves mIoU relative accuracy by 40.3 %. It is the first detailed study on object detection on fisheye cameras for autonomous driving scenarios to the best of our knowledge. The dataset comprising of 10,000 images along with all the object representations ground truth will be made public to encourage further research. We summarize our work in a short video with qualitative results at https://youtu.be/iLkOzvJpL-A .

****

A Pose Proposal and Refinement Network for Better 6D Object Pose Estimation

Ameni Trabelsi, Mohamed Chaabane, Nathaniel Blanchard, Ross Beveridge; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2382-2391

In this paper, we present a novel, end-to-end 6D object pose estimation method that operates on RGB inputs. Our approach is composed of 2 main components: the first component classifies the objects in the input image and proposes an initial 6D pose estimate through a multi-task, CNN-based encoder/multi-decoder module. The second component, a refinement module, includes a renderer and a multi-attentional pose refinement network, which iteratively refines the estimated poses by utilizing both appearance features and flow vectors. Our refiner takes advantage of the hybrid representation of the initial pose estimates to predict the relative errors with respect to the target poses. It is further augmented by a spatial multi-attention block that emphasizes objects' discriminative feature parts. Experiments on three benchmarks for 6D pose estimation show that our proposed pipeline outperforms state-of-the-art RGB-based methods with competitive runtime performance.

****

Meta Module Network for Compositional Visual Reasoning

Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, Jingjing Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 655-664

Neural Module Network (NMN) exhibits strong interpretability and compositionality thanks to its handcrafted neural modules with explicit multi-hop reasoning capability. However, most NMNs suffer from two critical drawbacks: 1) scalability: customized module for specific function renders it impractical when scaling up to a larger set of functions in complex tasks; 2) generalizability: rigid pre-defined module inventory makes it difficult to generalize to unseen functions in new tasks/domains. To design a more powerful NMN architecture for practical use, we propose Meta Module Network (MMN) centered on a novel meta module, which can take in function recipes and morph into diverse instance modules dynamically. The instance modules are then woven into an execution graph for complex visual reasoning, inheriting the strong explainability and compositionality of NMN. With such a flexible instantiation mechanism, the parameters of instance modules are inherited from the central meta module, retaining the same model complexity as the function set grows, which promises better scalability. Meanwhile, as functions are encoded into the embedding space, unseen functions can be readily represented based on its structural similarity with previously observed ones, which ensures better generalizability. Experiments on GQA and CLEVR datasets validate the superiority of MMN over state-of-the-art NMN designs. Synthetic experiments on held-out unseen functions from GQA dataset also demonstrate the strong generalizability of MMN.

****

Adaptiope: A Modern Benchmark for Unsupervised Domain Adaptation

Tobias Ringwald, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 101-110

Unsupervised domain adaptation (UDA) deals with the adaptation process of a given source domain with labeled training data to a target domain for which only unannotated data is available. This is a challenging task as the domain shift leads to degraded performance on the target domain data if not addressed. In this paper, we analyze commonly used UDA classification datasets and discover systematic problems with regard to dataset setup, ground truth ambiguity and annotation quality. We manually clean the most popular UDA dataset in the research area (Office-31) and quantify the negative effects of inaccurate annotations through thorough experiments. Based on these insights, we collect the Adaptiope dataset - a large scale, diverse UDA dataset with synthetic, product and real world data - and show that its transfer tasks provide a challenge even when considering recent UDA algorithms. Our datasets are available at https://gitlab.com/tringwald/adaptiope.
****

SinGAN-GIF: Learning a Generative Video Model From a Single GIF
Rajat Arora, Yong Jae Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1310-1319

We propose SinGAN-GIF, an extension of the image based SinGAN to GIFs or short video snippets. Our method learns the distribution of both the image patches in the GIF as well as their motion patterns. We do so by using a pyramid of 3D and 2D convolutional networks to model temporal information while reducing model parameters and training time, along with an image and a video discriminator. SinGAN-GIF can generate similar looking video samples for natural scenes at different spatial resolutions or temporal frame rates, and can be extended to other video applications like video editing, super resolution, and motion transfer. The project page, with supplementary video results, is: https://rajat95.github.io/singan-gif/
****

Minimal Solvers for Single-View Lens-Distorted Camera Auto-Calibration
Yaroslava Lochman, Oles Dobosevych, Rostyslav Hryniv, James Pritts; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2887-2896

This paper proposes minimal solvers that use combinations of imaged translational symmetries and parallel scene lines to jointly estimate lens undistortion with either affine rectification or focal length and absolute orientation. We use constraints provided by orthogonal scene planes to recover the focal length. We show that solvers using feature combinations can recover more accurate calibrations than solvers using only one feature type on scenes that have a balance of lines and texture. We also show that the proposed solvers are complementary and can be used together in a RANSAC-based estimator to improve auto-calibration accuracy. State-of-the-art performance is demonstrated on a standard dataset of lens-distorted urban images. The code is available at https://github.com/ylochman/single-view-autocalib.
****

Let's Get Dirty: GAN Based Data Augmentation for Camera Lens Soiling Detection in Autonomous Driving
Michal Uricar, Ganesh Sistu, Hazem Rashed, Antonin Vobecky, Varun Ravi Kumar, Pavel Krizek, Fabian Burger, Senthil Yogamani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 766-775

Wide-angle fisheye cameras are commonly used in automated driving for parking and low-speed navigation tasks. Four of such cameras form a surround-view system that provides a complete and detailed view of the vehicle. These cameras are directly exposed to harsh environmental settings and can get soiled very easily by mud, dust, water, frost. Soiling on the camera lens can severely degrade the visual perception algorithms, and a camera cleaning system triggered by a soiling detection algorithm is increasingly being deployed. While adverse weather conditions, such as rain, are getting attention recently, there is only limited work on general soiling. The main reason is the difficulty in collecting a diverse dataset as it is a relatively rare event. We propose a novel GAN based algorithm for generating unseen patterns of soiled images. Additionally, the proposed method a

utomatically provides the corresponding soiling masks eliminating the manual annotation cost. Augmentation of the generated soiled images for training improves the accuracy of soiling detection tasks significantly by 18% demonstrating its usefulness. The manually annotated soiling dataset and the generated augmentation dataset will be made public. We demonstrate the generalization of our fisheye trained GAN model on the Cityscapes dataset. We provide an empirical evaluation of the degradation of the semantic segmentation algorithm with the soiled data.
****

Single Image Human Proxemics Estimation for Visual Social Distancing

Maya Aghaei, Matteo Bustreo, Yiming Wang, Gianluca Bailo, Pietro Morerio, Alessio Del Bue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2785-2795

In this work, we address the problem of estimating the so-called "Social Distancing" given a single uncalibrated image in unconstrained scenarios. Our approach proposes a semi-automatic solution to approximate the homography matrix between the scene ground and image plane. With the estimated homography, we then leverage an off-the-shelf pose detector to detect body poses on the image and to reason upon their inter-personal distances using the length of their body-parts. Inter-personal distances are further locally inspected to detect possible violations of the social distancing rules. We validate our proposed method quantitatively and qualitatively against baselines on public domain datasets for which we provided groundtruth on inter-personal distances. Besides, we demonstrate the application of our method deployed in a real testing scenario where statistics on the inter-personal distances are currently used to improve the safety in a critical environment.
****

DeepMark++: Real-Time Clothing Detection at the Edge

Alexey Sidnev, Alexander Krapivin, Alexey Trushkov, Ekaterina Krasikova, Maxim Kazakov, Mikhail Viryasov; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2980-2988

Clothing recognition is the most fundamental AI application challenge within the fashion domain. While existing solutions offer decent recognition accuracy, they are generally slow and require significant computational resources. In this paper we propose a single-stage approach to overcome this obstacle and deliver rapid clothing detection and keypoint estimation. Our solution is based on a multi-target network CenterNet, and we introduce several powerful post-processing techniques to enhance performance. Our most accurate model achieves results comparable to state-of-the-art solutions on the DeepFashion2 dataset, and our light and fast model runs at 17 FPS on the Huawei P40 Pro smartphone. In addition, we achieved second place in the DeepFashion2 Landmark Estimation Challenge 2020 with 0.582 mAP on the test dataset.
****

Detecting Human-Object Interaction With Mixed Supervision

Suresh Kirthi Kumaraswamy, Miaojing Shi, Ewa Kijak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1228-1237

Human object interaction (HOI) detection is an important task in image understanding and reasoning. It is in a form of HOI triplet<human,verb,object> , requiring bounding boxes for humans and objects, and action be-tween them for the task completion. In other words, this task requires strong supervision for training, which is how-ever hard to procure. A natural solution to overcome this is to pursue weakly-supervised learning, where we only know the presence of certain HOI triplets in images but their ex-act location is unknown. Most weakly-supervised learning methods do not make provision for leveraging data with strong supervision, when they are available; and indeed a naive combination of this two paradigms in HOI detection fails to make contributions to each other. In this regard we propose a mixed-supervised HOI detection pipeline: thanks to a specific design of momentum-independent learning, it learns seamlessly across these two types of supervision. Moreover, in light of the annotation insufficiency in mixed supervision, we introduce an HOI element swap-ping technique to synthesize diverse and hard negatives across images and improve the robustness of the model. Our method i s evaluated on the challenging HICO-DET dataset. It outperforms the state of the art weakly- and fully-supervised methods under the same setting; and performs close to or even better than many fully-supervised methods by using a mixed amount of full and weak supervision.

****

Embedded Dense Camera Trajectories in Multi-Video Image Mosaics by Geodesic Interpolation-Based Reintegration

Lars Haalck, Benjamin Risse; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1849-1858

Dense registrations of huge image sets are still challenging due to exhaustive matchings and computationally expensive optimisations. Moreover, the resultant image mosaics often suffer from structural errors such as drift. Here, we propose a novel algorithm to generate global large-scale registrations from thousands of images extracted from multiple videos to derive high-resolution image mosaics which include full frame rate camera trajectories. Our algorithm does not require any initialisations and ensures the effective integration of all available image data by combining efficient and highly parallelised key-frame and loop-closure mechanisms with a novel geodesic interpolation-based reintegration strategy. As a consequence, global refinement can be done in a fraction of iterations compared to traditional optimisation strategies, while effectively avoiding drift and convergence towards inappropriate solutions. We compared our registration strategy with state-of-the-art algorithms and quantitative evaluations revealed millimetre spatial and high angular accuracy. Applicability is demonstrated by registering more than 110,000 frames from multiple scan recordings and provide dense camera trajectories in a globally referenced coordinate system as used for drone-based mappings, ecological studies, object tracking and land surveys.

****

Self Supervision for Attention Networks

Badri N. Patro, Kasturi G.S., Ansh Jain, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 726-735

In recent years, the attention mechanism has become a fairly popular concept and has proven to be successful in many machine learning applications. However, deep learning models do not employ supervision for these attention mechanisms which can improve the model's performance significantly. Therefore, in this paper, we tackle this limitation and propose a novel method to improve the attention mechanism by inducing "self-supervision". We devise a technique to generate desirable attention maps for any model that utilizes an attention module. This is achieved by examining the model's output for different regions sampled from the input and obtaining the attention probability distributions that enhance the proficiency of the model. The attention distributions thus obtained are used for supervision. We rely on the fact, that attenuation of the unimportant parts, allows a model to attend to more salient regions, thus strengthening the prediction accuracy. The quantitative and qualitative results published in this paper show that this method successfully improves the attention mechanism as well as the model's accuracy. In addition to the task of Visual Question Answering(VQA), we also show results on the task of Image classification and Text classification to prove that our method can be generalized to any vision and language model that uses an attention module

****

QuadroNet: Multi-Task Learning for Real-Time Semantic Depth Aware Instance Segmentation

Kratarth Goel, Praveen Srinivasan, Sarah Tariq, James Philbin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 315-324

Vision for autonomous driving is a uniquely challenging problem: the number of tasks required for full scene understanding is large and diverse; the quality requirements on each task are stringent due to the safety-critical nature of the application; and the latency budget is limited, requiring real-time solutions. In this work we address these challenges with QuadroNet, a one-shot network that jo intly produces four outputs: 2D detections, instance segmentation, semantic segmentation, and monocular depth estimates in real-time (>60fps) on consumer-grade GPU hardware. On a challenging real-world autonomous driving dataset, we demonstrate an increase of +2.4% mAP for detection, +3.15% mIoU for semantic segmentation, +5.05% mAP@0.5 for instance segmentation and +1.36% in delta<1.25 for depth prediction over a baseline approach. We also compare our work against other multi-task learning approaches on Cityscapes and demonstrate state-of-the-art results.

****

3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-View Model-Fitting

Zhongguo Li, Magnus Oskarsson, Anders Heyden; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1888-1897

3D human pose and shape estimation plays a vital role in many computer vision applications. There are many deep learning based methods attempting to solve the problem only relying on single-view RGB images for training the network. However, since some public datasets are captured from multi-view cameras system, we propose a novel method to tackle the problem by putting optimization-based multi-view model-fitting into a regression-based learning loop from multi-view images. Firstly, a convolutional neural network (CNN) regresses the pose and shape of a parametric human body model (SMPL) from multi-view images. Then, utilizing the regressed pose and shape as initialization, we propose an improved multi-view optimization method based on the SMPLify method (MV-SMPLify) to fit the SMPL model to the multi-view images simultaneously. Subsequently, the optimized parameters can be adopted to supervise the training of the CNN model. This whole process forms a self-supervising framework which can combine the advantages of the CNN approach and the optimization-based approach through a collaborative process. In addition, the multi-view images can provide more comprehensive supervision for the training. Experiments on public datasets qualitatively and quantitatively demonstrate that our method outperforms previous approaches in a number of ways.

****

Temporal Shift GAN for Large Scale Video Generation

Andres Munoz, Mohammadreza Zolfaghari, Max Argus, Thomas Brox; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3179-3188

Video generation models have become increasingly popular in the last few years, however the standard 2D architectures used today lack natural spatio-temporal modelling capabilities. In this paper, we present a network architecture for video generation that models spatio-temporal consistency without resorting to costly 3D architectures. The architecture facilitates information exchange between neighboring time points, which improves the temporal consistency of both the high level structure as well as the low-level details of the generated frames. The approach achieves state-of-the-art quantitative performance, as measured by the inception score on the UCF-101 dataset as well as better qualitative results. We also introduce a new quantitative measure (S3) that uses downstream tasks for evaluation. Moreover, we present a new multi-label dataset MaisToy, which enables us to evaluate the generalization of the model.

****

Only Time Can Tell: Discovering Temporal Data for Temporal Modeling

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, Lorenzo Torresani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 535-544

Understanding temporal information and how the visual world changes over time, is a fundamental ability of intelligent systems. In video understanding, temporal information is at the core of many current challenges, including compression, efficient inference, motion estimation or summarization. However, in current video datasets it has been observed that action classes can often be recognized without any temporal information, from a single frame of video. As a result, both benchmarking and training in these datasets may give an unintentional advantage to models with strong image understanding capabilities, as opposed to those with s trong temporal understanding, potentially hindering progress. In this paper we address this problem head on by identifying action classes where temporal information is actually necessary to recognize them and call these "temporal classes". Selecting temporal classes using a computational method would bias the process. Instead, we propose a methodology based on a simple and effective human annotation experiment. We remove just the temporal information, by shuffling frames in time, and measure if the action can still be recognized. Classes that cannot be recognized when frames are not in order, are included in the temporal set. We observe that this set is statistically different from other static classes, and that performance in it correlates with a network's ability to capture temporal information. Thus we use it as a benchmark on current popular networks, which reveals a series of interesting facts, like inflated convolutions bias networks towards classes where motion is not important. We also explore the effect of training on the temporal set, and observe that this leads to better generalization in unseen classes, demonstrating the need for more temporal data. We hope that the proposed dataset of temporal categories will help guide future research in temporal modeling for better video understanding.

****

Data-Efficient Alignment of Multimodal Sequences by Aligning Gradient Updates and Internal Feature Distributions

Jianan Wang, Boyang Li, Xiangyu Fan, Jing Lin, Yanwei Fu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 665-675

The task of video and text sequence alignment is a prerequisite step toward joint understanding of movie videos and screenplays. However, supervised methods face the obstacle of limited realistic training data. With this paper, we attempt to enhance data efficiency of the end-to-end alignment network NeuMATCH [15]. Recent research [56] suggests that network components dealing with different modalities may overfit and generalize at different speeds, creating difficulties for training. We propose to employ (1) layer-wise adaptive rate scaling (LARS) to align the magnitudes of gradient updates in different layers and balance the pace of learning and (2) sequence-wise batch normalization (SBN) to align the internal feature distributions from different modalities. Finally, we leverage random projection to reduce the dimensionality of input features. On the YouTube Movie Summary dataset, the combined use of these technique closes the performance gap when the pretraining on the LSMDC dataset is omitted and achieves the state-of-the-art result. Extensive empirical comparisons and analysis reveal that these techniques improve optimization and regularize the network more effectively than two different setups of layer normalization.

****

Are These From the Same Place? Seeing the Unseen in Cross-View Image Geo-Localization

Royston Rodrigues, Masahiro Tani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3753-3761

In an era where digital maps act as gateways to exploring the world, the availability of large scale geo-tagged imagery has inspired a number of visual navigation techniques. One promising approach to visual navigation is cross-view image geo-localization. Here, the images whose location needs to be determined are matched against a database of geo-tagged aerial imagery. The methods based on this approach sought to resolve view point changes. But scenes also vary temporally, during which new landmarks might appear or existing ones might disappear. One cannot guarantee storage of aerial imagery across all time instants and hence a technique robust to temporal variation in scenes becomes of paramount importance. In this paper, we address the temporal gap between scenes by proposing a two step approach. First, we propose a semantically driven data augmentation technique that gives Siamese networks the ability to hallucinate unseen objects. Then we present the augmented samples to a multi-scale attentive embedding network to perform matching tasks. Experiments on standard benchmarks demonstrate the integration of the proposed approach with existing frameworks improves top-1 image recall rate on the CVUSA data-set from 89.84 % to 93.09 %, and from 81.03 % to 87.21 %

on the CVACT data-set.

****

DualSR: Zero-Shot Dual Learning for Real-World Super-Resolution

Mohammad Emad, Maurice Peemen, Henk Corporaal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1630-1639

Advanced methods for single image super-resolution (SISR) based upon Deep learning have demonstrated a remarkable reconstruction performance on downscaled images. However, for real-world low-resolution images (e.g. images captured straight from the camera) they often generate blurry images and highlight unpleasant artifacts. The main reason is the training data that does not reflect the real-world super-resolution problem. They train the network using images downsampled with an ideal (usually bicubic) kernel. However, for real-world images the degradation process is more complex and can vary from image to image. This paper proposes a new dual-path architecture (DualSR) that learns an image-specific low-to-high resolution mapping using only patches of the input test image. For every image, a downsampler learns the degradation process using a generative adversarial network, and an upsampler learns to super-resolve that specific image. In the DualSR architecture, the upsampler and downsampler are trained simultaneously and they improve each other using cycle consistency losses. For better visual quality and eliminating undesired artifacts, the upsampler is constrained by a masked interpolation loss. On standard benchmarks with unknown degradation kernels, DualSR outperforms recent blind and non-blind super-resolution methods in term of SSIM and generates images with higher perceptual quality. On real-world LR images it generates visually pleasing and artifact-free results.

****

Subsurface Pipes Detection Using DNN-Based Back Projection on GPR Data

Jinglun Feng, Liang Yang, Haiyan Wang, Yingli Tian, Jizhong Xiao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 266-275

Localization and reconstruction of underground targets, the problem of estimating the position and geometry of the objects from Ground Penetration Radar (GPR), still lies at the core of non-destructive testing (NDT). In this paper, we present MigrationNet, a learning-based approach to detect and visualize subsurface objects. Compared with the existing learning-based method of GPR, our proposed approach could not only detect the hyperbola feature in the raw B-scan image but also interpret hyperbola features into the cross-section image of subsurface pipes. Furthermore, to compare the proposed method with the conventional back-projection methods for GPR data interpretation, a synthetic GPR dataset that mimics the real NDT environment is also introduced in this work. The study indicates the effectiveness of our method, it uses less GPR data for underground pipes reconstruction, produces better GPR imaging results with less computation, and shows the robustness to noise.

****

Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach

Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, Barbara Caputo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2918-2927

We address the task of cross-domain visual place recognition, where the goal is to geolocalize a given query image against a labeled gallery, in the case where the query and the gallery belong to different visual domains. To achieve this, we focus on building a domain robust deep network by leveraging over an attention mechanism combined with few-shot unsupervised domain adaptation techniques, where we use a small number of unlabeled target domain images to learn about the target distribution. With our method, we are able to outperform the current state of the art while using two orders of magnitude less target domain images. Finally we propose a new large-scale dataset for cross-domain visual place recognition, called SVOX. The pytorch code is available at https://github.com/valeriopaolicelli/AdAGeo .

****

S-VVAD: Visual Voice Activity Detection by Motion Segmentation

Muhammad Shahid, Cigdem Beyan, Vittorio Murino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2332-2341

We address the challenging Voice Activity Detection (VAD) problem, which determines "Who is Speaking and When?" in audiovisual recordings. The typical audio-based VAD systems can be ineffective in the presence of ambient noise or noise variations. Moreover, due to technical or privacy reasons, audio might not be always available. In such cases, the use of video modality to perform VAD is desirable. Almost all existing visual VAD methods rely on body part detection, e.g., face, lips, or hands. In contrast, we propose a novel visual VAD method operating directly on the entire video frame, without the explicit need of detecting a person or his/her body parts. Our method, named S-VVAD, learns body motion cues associated with speech activity within a weakly supervised segmentation framework. Therefore, it not only detects the speakers/not-speakers but simultaneously localizes the image positions of them. It is an end-to-end pipeline, person-independent and it does not require any prior knowledge nor pre-processing. S-VVAD performs well in various challenging conditions and demonstrates the state-of-the-art results on multiple datasets. Moreover, the better generalization capability of S-VVAD is confirmed for cross-dataset and person-independent scenarios.

****

Deep Poisoning: Towards Robust Image Data Sharing Against Visual Disclosure

Hao Guo, Brian Dolhansky, Eric Hsin, Phong Dinh, Cristian Canton Ferrer, Song Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 686-696

Due to respectively limited training data, different entities addressing the same vision task based on certain sensitive images may not train a robust deep network. This paper introduces a new vision task where various entities share task-specific image data to enlarge each other's training data volume without visually disclosing sensitive contents (e.g. illegal images). Then, we present a new structure-based training regime to enable different entities learn task-specific and reconstruction-proof image representations for image data sharing. Specifically, each entity learns a private Deep Poisoning Module (DPM) and insert it to a pre-trained deep network, which is designed to perform the specific vision task. The DPM deliberately poisons convolutional image features to prevent image reconstructions, while ensuring that the altered image data is functionally equivalent to the non-poisoned data for the specific vision task. Given this equivalence, the poisoned features shared from one entity could be used by another entity for further model refinement. Experimental results on image classification prove the efficacy of the proposed method.

****

RODNet: Radar Object Detection Using Cross-Modal Supervision

Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, Hui Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 504-513

Radar is usually more robust than the camera in severe driving scenarios, e.g., weak/strong lighting and bad weather. However, unlike RGB images captured by a camera, the semantic information from the radar signals is noticeably difficult to extract. In this paper, we propose a deep radar object detection network (RODNet), to effectively detect objects purely from the carefully processed radar frequency data in the format of range-azimuth frequency heatmaps (RAMaps). Three different 3D autoencoder based architectures are introduced to predict object confidence distribution from each snippet of the input RAMaps. The final detection results are then calculated using our post-processing method, called location-based non-maximum suppression (L-NMS). Instead of using burdensome human-labeled ground truth, we train the RODNet using the annotations generated automatically by a novel 3D localization method using a camera-radar fusion (CRF) strategy. To train and evaluate our method, we build a new dataset -- CRUW, containing synchronized videos and RAMaps in various driving scenarios. After intensive experiments, our RODNet shows favorable object detection performance without the presence of the camera.

****

Assessing Image and Text Generation With Topological Analysis and Fuzzy Logic
Goncalo Mordido, Julian Niedermeier, Christoph Meinel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2013-2022
Objective and interpretable metrics to evaluate current artificial intelligent systems are of great importance, not only to analyze the current state of such systems but also to objectively measure progress in the future. We propose a novel metric, called Fuzzy Topology Impact (FTI), that assesses both the quality and diversity of a generated set using topological representations combined with fuzzy logic. In our synthetic experiments, FTI consistently outperforms current evaluation methods in terms of stability and sensitivity to detect drops in quality and diversity in the generated set, both on image and text generation tasks. Moreover, FTI shows a high degree of correlation to human evaluation on unconditional language generation.
****
R-MNet: A Perceptual Adversarial Network for Image Inpainting
Jireh Jam, Connah Kendrick, Vincent Drouard, Kevin Walker, Gee-Sern Hsu, Moi Hoon Yap; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2714-2723
Facial image inpainting is a problem that is widely studied, and in recent years the introduction of Generative Adversarial Networks, has led to improvements in the field. Unfortunately some issues persists, in particular when blending the missing pixels with the visible ones. We address the problem by proposing a Wasserstein GAN combined with a new reverse mask operator, namely Reverse Masking Network (R-MNet), a perceptual adversarial network for image inpainting. The reverse mask operator transfers the reverse masked image to the end of the encoder-decoder network leaving only valid pixels to be inpainted. Additionally, we propose a new loss function computed in feature space to target only valid pixels combined with adversarial training. These then capture data distributions and generate images similar to those in the training data with achieved realism (realistic and coherent) on the output images. We evaluate our method on publicly available dataset, and compare with state-of-the-art methods. We show that our method is able to generalize to high-resolution inpainting task, and further show more realistic outputs that are plausible to the human visual system when compared with the state-of-the-art methods. https://github.com/Jireh-Jam/R-MNET-A-Perceptual-Adversarial-Network-for-Image-Inpainting
****
Representation Learning With Statistical Independence to Mitigate Bias
Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, Kilian M. Pohl; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2513-2523
Presence of bias (in datasets or tasks) is inarguably one of the most critical challenges in machine learning applications that has alluded to pivotal debates in recent years. Such challenges range from spurious associations between variables in medical studies to the bias of race in gender or face recognition systems. Controlling for all types of biases in the dataset curation stage is cumbersome and sometimes impossible. The alternative is to use the available data and build models incorporating fair representation learning. In this paper, we propose such a model based on adversarial training with two competing objectives to learn features that have (1) maximum discriminative power with respect to the task and (2) minimal statistical mean dependence with the protected (bias) variable(s). Our approach does so by incorporating a new adversarial loss function that encourages a vanished correlation between the bias and the learned features. We apply our method to synthetic data, medical images (containing task bias), and a dataset for gender classification (containing dataset bias). Our results show that the learned features by our method not only result in superior prediction performance but also are unbiased. The code is available at https:// github.com/QingyuZhao/BR-Net/ .
****
G2D: Generate to Detect Anomaly

Masoud Pourreza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, Mohammad Sabokrou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2003-2012

In this paper, we propose a novel method for irregularity detection. Previous researches solve this problem as a One-Class Classification (OCC) task where they train a reference model on all of the available samples. Then, they consider a test sample as an anomaly if it has a diversion from the reference model. Generative Adversarial Networks (GANs) have achieved the most promising results for OCC while implementing and training such networks, especially for the OCC task, is a cumbersome and computationally expensive procedure. To cope with the mentioned challenges, we present a simple but effective method to solve the irregularity detection as a binary classification task in order to make the implementation easier along with improving the detection performance. We learn two deep neural networks (generator and discriminator) in a GAN-style setting on merely the normal samples. During training, the generator gradually becomes an expert to generate samples which are similar to the normal ones. In the training phase, when the generator fails to produce normal data (in the early stages of learning and also prior to the complete convergence), it can be considered as an irregularity generator. In this way, we simultaneously generate the irregular samples. Afterward, we train a binary classifier on the generated anomalous samples along with the normal instances in order to be capable of detecting irregularities. The proposed framework applies to different related applications of outlier and anomaly detection in images and videos, respectively. The results confirm that our proposed method is superior to the baseline and state-of-the-art solutions.

****

Compositional Embeddings for Multi-Label One-Shot Learning

Zeqian Li, Michael Mozer, Jacob Whitehill; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 296-304

We present a compositional embedding framework that infers not just a single class per input image, but a set of classes, in the setting of one-shot learning. Specifically, we propose and evaluate several novel models consisting of (1) an embedding function f trained jointly with a "composition" function g that computes set union operations between the classes encoded in two embedding vectors; and (2) embedding f trained jointly with a "query" function h that computes whether the classes encoded in one embedding subsume the classes encoded in another embedding. In contrast to prior work, these models must both perceive the classes associated with the input examples and encode the relationships between different class label sets, and they are trained using only weak one-shot supervision consisting of the label-set relationships among training examples. Experiments on the OmniGlot, Open Images, and COCO datasets show that the proposed compositional embedding models outperform existing embedding methods. Our compositional embedding models have applications to multi-label object recognition for both one-shot and supervised learning.

****

Focus and Retain: Complement the Broken Pose in Human Image Synthesis

Pu Ge, Qiushi Huang, Wei Xiang, Xue Jing, Yule Li, Yiyong Li, Zhun Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3370-3379

Given a target pose, how to generate an image of a specific style with that target pose remains an ill-posed and thus complicated problem. Most recent works treat the human pose synthesis tasks as an image spatial transformation problem using flow warping techniques. However, we observe that, due to the inherent ill-posed nature of many complicated human poses, former methods fail to generate body parts. To tackle this problem, we propose a feature-level flow attention module and an Enhancer Network. The flow attention module produces a flow attention mask to guide the combination of the flow-warped features and the structural pose features. Then, we apply the Enhancer Network to refine the coarse image by injecting the pose information. We present our experimental evaluation both qualitatively and quantitatively on DeepFashion, Market-1501, and Youtube dance datasets. Quantitative results show that our method has 12.995 FID at DeepFashion, 25.45

9 FID at Market-1501, 14.516 FID at Youtube dance datasets, which outperforms some state-of-the-arts including Guide-Pixe2Pixe, Global-Flow-Local-Attn, and CocosNet.

****

MeliusNet: An Improved Network Architecture for Binary Neural Networks

Joseph Bethge, Christian Bartz, Haojin Yang, Ying Chen, Christoph Meinel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1439-1448

Binary Neural Networks (BNNs) are neural networks which use binary weights and activations instead of the typical 32-bit floating point values. They have reduced model sizes and allow for efficient inference on mobile or embedded devices with limited power and computational resources. However, the binarization of weights and activations leads to feature maps of lower quality and lower capacity and thus a drop in accuracy compared to their 32-bit counterparts. Previous work has increased the number of channels or used multiple binary bases to alleviate these problems. In this paper, we instead present an architectural approach: MeliusNet. It consists of alternating a DenseBlock, which increases the feature capacity, and our proposed ImprovementBlock, which increases the feature quality. Experiments on the ImageNet dataset demonstrate the superior performance of our MeliusNet over a variety of popular binary architectures with regards to both computation savings and accuracy. Furthermore, BNN models trained with our method can match the accuracy of the popular compact network MobileNet-v1 in terms of model size and number of operations.

****

Facial Expression Recognition in the Wild via Deep Attentive Center Loss

Amir Hossein Farzaneh, Xiaojun Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2402-2411

Learning discriminative features for Facial Expression Recognition (FER) in the wild using Convolutional Neural Networks (CNNs) is a non-trivial task due to the significant intra-class variations and inter-class similarities. Deep Metric Learning (DML) approaches such as center loss and its variants jointly optimized with softmax loss have been adopted in many FER methods to enhance the discriminative power of learned features in the embedding space. However, equally supervising all features with the metric learning method might include irrelevant features and ultimately degrade the generalization ability of the learning algorithm. We propose a Deep Attentive Center Loss (DACL) method to adaptively select a subset of significant feature elements for enhanced discrimination. The proposed DACL integrates an attention mechanism to estimate attention weights correlated with feature importance using the intermediate spatial feature maps in CNN as context. The estimated weights accommodate the sparse formulation of center loss to selectively achieve intra-class compactness and inter-class separation for the relevant information in the embedding space. An extensive study on two widely used wild FER datasets demonstrates the superiority of the proposed DACL method compared to state-of-the-art methods.

****

Attentional Feature Fusion

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, Kobus Barnard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3560-3569

Feature fusion, the combination of features from different layers or branches, is an omnipresent part of modern network architectures. It is often implemented via simple operations, such as summation or concatenation, but this might not be the best choice. In this work, we propose a uniform and general scheme, namely attentional feature fusion, which is applicable for most common scenarios, including feature fusion induced by short and long skip connections as well as within Inception layers. To better fuse features of inconsistent semantics and scales, we propose a multi-scale channel attention module, which addresses issues that arise when fusing features given at different scales. We also demonstrate that the initial integration of feature maps can become a bottleneck and that this issue can be alleviated by adding another level of attention, which we refer to as i terative attentional feature fusion. With fewer layers or parameters, our models outperform state-of-the-art networks on both CIFAR-100 and ImageNet datasets, which suggests that more sophisticated attention mechanisms for feature fusion hold great potential to consistently yield better results compared to their direct counterparts. Our codes and trained models are available online.

****

DeepCFL: Deep Contextual Features Learning From a Single Image
Indra Deep Mastan, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2897-2906

Recently, there is a vast interest in developing image feature learning methods that are independent of the training data, such as deep image prior, InGAN, SinGAN, and DCIL. These methods are unsupervised and are used to perform low-level vision tasks such as image restoration, image editing, and image synthesis. In this work, we proposed a new training data-independent framework, called Deep Contextual Features Learning (DeepCFL), to perform image synthesis and image restoration based on the semantics of the input image. The contextual features are simply the high dimensional vectors representing the semantics of the given image. DeepCFL is a single image GAN framework that learns the distribution of the context vectors from the input image. We show the performance of contextual learning in various challenging scenarios: outpainting, inpainting, and restoration of randomly removed pixels. DeepCFL is applicable when the input source image and the generated target image are not aligned. We illustrate image synthesis using DeepCFL for the task of image resizing.

****

Accelerated WGAN Update Strategy With Loss Change Rate Balancing
Xu Ouyang, Ying Chen, Gady Agam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2546-2555

Optimizing the discriminator in Generative Adversarial Networks (GANs) to completion in the inner training loop is computationally prohibitive, and on finite datasets would result in overfitting. To address this, a common update strategy is to alternate between k optimization steps for the discriminator D and one optimization step for the generator G. This strategy is repeated in various GAN algorithms where k is selected empirically. In this paper, we show that this update strategy is not optimal in terms of accuracy and convergence speed, and propose a new update strategy for networks with Wasserstein GAN (WGAN) group related loss functions (e.g. WGAN, WGAN-GP, Deblur GAN, and Super resolution GAN). The proposed update strategy is based on a loss change ratio comparison of G and D. We demonstrate that the proposed strategy improves both convergence speed and accuracy.

****

End-to-End Chinese Landscape Painting Creation Using Generative Adversarial Networks
Alice Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3863-3871

Current GAN-based art generation methods produce unoriginal artwork due to their dependence on conditional input. Here, we propose Sketch-And-Paint GAN (SAPGAN), the first model which generates Chinese landscape paintings from end to end, without conditional input. SAPGAN is composed of two GANs: SketchGAN for generation of edge maps, and PaintGAN for subsequent edge-to-painting translation. Our model is trained on a new dataset of traditional Chinese landscape paintings never before used for generative research. A 242-person Visual Turing Test study reveals that SAPGAN paintings are mistaken as human artwork with 55% frequency, significantly outperforming paintings from baseline GANs. Our work lays a groundwork for truly machine-original art generation.

****

Coarse Temporal Attention Network (CTA-Net) for Driver's Activity Recognition
Zachary Wharton, Ardhendu Behera, Yonghuai Liu, Nik Bessis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1279-1289

There is significant progress in recognizing traditional human activities from v ideos focusing on highly distinctive actions involving discriminative body movements, body-object and/or human-human interactions. Driver's activities are different since they are executed by the same subject with similar body parts movements, resulting in subtle changes. To address this, we propose a novel framework by exploiting the spatiotemporal attention to model the subtle changes. Our model is named Coarse Temporal Attention Network (CTA-Net), in which coarse temporal branches are introduced in a trainable glimpse network. The goal is to allow the glimpse to capture high-level temporal relationships, such as 'during', 'before' and 'after' by focusing on a specific part of a video. These branches also respect the topology of the temporal dynamics in the video, ensuring that different branches learn meaningful spatial and temporal changes. The model then uses an innovative attention mechanism to generate high-level action specific contextual information for activity recognition by exploring the hidden states of an LSTM. The attention mechanism helps in learning to decide the importance of each hidden state for the recognition task by weighing them when constructing the representation of the video. Our approach is evaluated on four publicly accessible datasets and significantly outperforms the state-of-the-art by a considerable margin with only RGB video as input.

****

Defense-Friendly Images in Adversarial Attacks: Dataset and Metrics for Perturbation Difficulty

Camilo Pestana, Wei Liu, David Glance, Ajmal Mian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 556-565

Dataset bias is a problem in adversarial machine learning, especially in the evaluation of defenses. An adversarial attack or defense algorithm may show better results on the reported dataset than can be replicated on other datasets. Even when two algorithms are compared, their relative performance can vary depending on the dataset. Deep learning offers state-of-the-art solutions for image recognition, but deep models are vulnerable even to small perturbations. Research in this area focuses primarily on adversarial attacks and defense algorithms. In this paper, we report for the first time, a class of robust images that are both resilient to attacks and that recover better than random images under adversarial attacks using simple defense techniques. Thus, a test dataset with a high proportion of robust images gives a misleading impression about the performance of an adversarial attack or defense. We propose three metrics to determine the proportion of robust images in a dataset and provide scoring to determine the dataset bias. We also provide an ImageNet-R dataset of 15000+ robust images to facilitate further research on this intriguing phenomenon of image strength under attack. Our dataset, combined with the proposed metrics, is valuable for unbiased benchmarking of adversarial attack and defense algorithms.

****

DANCE: A Deep Attentive Contour Model for Efficient Instance Segmentation

Zichen Liu, Jun Hao Liew, Xiangyu Chen, Jiashi Feng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 345-354

Contour-based instance segmentation methods are attractive due to their efficiency. However, existing contour-based methods either suffer from lossy representation, complex pipeline or difficulty in model training, resulting in subpar mask accuracy on challenging datasets like MS-COCO. In this work, we propose a novel deep attentive contour model, named DANCE, to achieve better instance segmentation accuracy while remaining good efficiency. To this end, DANCE applies two new designs: attentive contour deformation to refine the quality of segmentation contours and segment-wise matching to ease the model training. Comprehensive experiments demonstrate DANCE excels at deforming the initial contour in a more natural and efficient way towards the real object boundaries. Effectiveness of DANCE is also validated on the COCO dataset, which achieves 38.1% mAP and outperforms all other contour-based instance segmentation models. To the best of our knowledge, DANCE is the first contour-based model that achieves comparable performance to pixel-wise segmentation models. Code is available at https://github.com/lkevinzc/dance.

****

Holistic Filter Pruning for Efficient Deep Neural Networks
Lukas Enderich, Fabian Timm, Wolfram Burgard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2596-2605
Deep neural networks (DNNs) are usually over-parameterized to increase the likelihood of getting adequate initial weights by random initialization. Consequently, trained DNNs have many redundancies which can be pruned from the model to reduce complexity and improve the ability to generalize. Structural sparsity, as achieved by filter pruning, directly reduces the tensor sizes of weights and activations and is thus particularly effective for reducing complexity. We propose Holistic Filter Pruning (HFP), a novel approach for common DNN training that is easy to implement and enables to specify accurate pruning rates for the number of both parameters and multiplications. After each forward pass, the current model size is calculated and compared to the desired target size. By gradient descent, a global solution can be found that allocates the pruning budget over the individual layers such that the desired target size is fulfilled. In various experiments, we give insights into the training and achieve state-of-the-art performance on CIFAR-10 and ImageNet.
****
Learning to Generate Dense Point Clouds With Textures on Multiple Categories
Tao Hu, Geng Lin, Zhizhong Han, Matthias Zwicker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2170-2179
3D reconstruction from images is a core problem in computer vision. With recent advances in deep learning, it has become possible to recover plausible 3D shapes even from single RGB images. However, obtaining detailed geometry and texture for objects with arbitrary topology remains challenging. In this paper, we propose a novel approach for reconstructing point clouds from RGB images. Unlike other methods, we can recover dense point clouds with hundreds of thousands of points, and we also include RGB textures. In addition, we train our model on multiple categories, which leads to superior generalization to unseen categories compared to previous techniques. We achieve this using a two-stage approach, where we first infer an object coordinate map from the input RGB image, and then obtain the final point cloud using a reprojection and completion step. We show results on standard benchmarks that demonstrate the advantages of our technique.
****
Same Same but DifferNet: Semi-Supervised Defect Detection With Normalizing Flows
Marco Rudolph, Bastian Wandt, Bodo Rosenhahn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1907-1916
The detection of manufacturing errors is crucial in fabrication processes to ensure product quality and safety standards. Since many defects occur very rarely and their characteristics are mostly unknown a priori, their detection is still an open research question. To this end, we propose DifferNet: It leverages the descriptiveness of features extracted by convolutional neural networks to estimate their density using normalizing flows. Normalizing flows are well-suited to deal with low dimensional data distributions. However, they struggle with the high dimensionality of images. Therefore, we employ a multi-scale feature extractor which enables the normalizing flow to assign meaningful likelihoods to the images. Based on these likelihoods we develop a scoring function that indicates defects. Moreover, propagating the score back to the image enables pixel-wise localization. To achieve a high robustness and performance we exploit multiple transformations in training and evaluation. In contrast to most other methods, ours does not require a large number of training samples and performs well with as low as 16 images. We demonstrate the superior performance over existing approaches on the challenging and newly proposed MVTec AD and Magnetic Tile Defects datasets.
****
Weakly-Supervised Object Representation Learning for Few-Shot Semantic Segmentation
Xiaowen Ying, Xin Li, Mooi Choo Chuah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1497-1506
Training a semantic segmentation model requires large densely-annotated image datasets that are costly to obtain. Once the training is done, it is also difficul

t to add new object categories to such segmentation models. In this paper, we tackle the few-shot semantic segmentation problem, which aims to perform image segmentation task on unseen object categories merely based on one or a few support example(s). The key to solving this few-shot segmentation problem lies in effectively utilizing object information from support examples to separate target objects from the background in a query image. While existing methods typically generate object-level representations by averaging local features in support images, we demonstrate that such object representations are typically noisy and less distinguishing. To solve this problem, we design an object representation generator (ORG) module which can effectively aggregate local object features from support image(s) and produce better object-level representation. The ORG module can be embedded into the network and trained end-to-end in a weakly-supervised fashion without extra human annotation. We incorporate this design into a modified encoder-decoder network to present a powerful and efficient framework for few-shot semantic segmentation. Experimental results on the Pascal-VOC and MS-COCO datasets show that our approach achieves better performance compared to existing methods under both one-shot and five-shot settings.

****

SubICap: Towards Subword-Informed Image Captioning

Naeha Sharif, Mohammed Bennamoun, Wei Liu, Syed Afaq Ali Shah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3540-3549

Existing Image Captioning (IC) systems model words as atomic units in captions and are unable to exploit the structural information in the words. This makes representation of rare words very difficult and out-of-vocabulary words impossible. Moreover, to avoid computational complexity, existing IC models operate over a modest sized vocabulary of frequent words, such that the identity of rare words is lost. In this work we address this common limitation of IC systems in dealing with rare words in the corpora. We decompose words into smaller constituent units `subwords' and represent captions as a sequence of subwords instead of words. This helps represent all words in the corpora using a significantly lower subword vocabulary, leading to better parameter learning. Using subword language modeling, our captioning system improves various metric scores, with a training vocabulary size approximately 90% less than the baseline and various state-of-the-art word-level models. Our quantitative and qualitative results and analysis signify the efficacy of our proposed approach.

****

Faces a la Carte: Text-to-Face Generation via Attribute Disentanglement

Tianren Wang, Teng Zhang, Brian Lovell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3380-3388

Text-to-Face (TTF) synthesis is a challenging task with great potential for diverse computer vision applications. Compared to Text-to-Image (TTI) synthesis tasks, the textual description of faces can be much more complicated and detailed due to the variety of facial attributes and the parsing of high dimensional abstract natural language. In this paper, we propose a Text-to-Face model that not only produces images in high resolution (1024*1024) with text-to-image consistency, but also outputs multiple diverse faces to cover a wide range of unspecified facial features in a natural way. By fine-tuning the multi-label classifier and image encoder, our model obtains the adjustment vectors and image embeddings which are used to transform the input noise vector sampled from the normal distribution. Afterwards, the transformed noise vector is fed into a pre-trained high-resolution image generator to produce a set of faces with the desired facial attributes. We refer to our model as TTF-HD. Experimental results show that TTF-HD generates high-quality synthesised faces from free-

****

Scaling Digital Screen Reading With One-Shot Learning and Re-Identification

James Charles, Stefano Bucciarelli, Roberto Cipolla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2635-2643

Using only a mobile phone app, our objective is to cheaply retro-fit digital met ers (e.g blood pressure, blood glucose or industrial gauges) with `smart' data transfer capabilities. Using the mobile phone camera we build an app to securely and accurately transcribe information from digital meter screens. Only a single labelled training image of a target meter is required to build a custom screen reading module. Here we show how this can scale to potentially hundreds of different meters by learning to recognising the meter type so that the reading module can be automatically selected. This makes the system very easy for a user who would need to scan multiple different meter types. To this end, we build a CNN based system which runs in real-time on mobile device with very high read accuracy and meter recognition. Our contributions include (i) a method of one-shot training by synthesis through domain shift reduction, (ii) a deep embedding network for scale, translation and rotation invariant re-identification of digital meters, (iii) a highly accurate and efficient mobile phone app for recognising and parsing digital meter screens and (iv) release of a new digital meter re-identification dataset.

****

Unsupervised Attention Based Instance Discriminative Learning for Person Re-Identification

Kshitij Nikhal, Benjamin S. Riggan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2422-2431

Recent advances in person re-identification have demonstrated enhanced discriminability, especially with supervised learning or transfer learning. However, since the data requirements---including the degree of data curations---are becoming increasingly complex and laborious, there is a critical need for unsupervised methods that are robust to large intra-class variations, such as changes in perspective, illumination, articulated motion, resolution, etc. Therefore, we propose an unsupervised framework for person re-identification which is trained in an end-to-end manner without any pre-training. Our proposed framework leverages a new attention mechanism that combines group convolutions to (1) enhance spatial attention at multiple scales and (2) reduce the number of trainable parameters by 59.6%. Additionally, our framework jointly optimizes the network with agglomerative clustering and instance learning to tackle hard samples. We perform extensive analysis using the Market1501 and DukeMTMC-reID datasets to demonstrate that our method consistently outperforms the state-of-the-art methods (with and without pre-trained weights).

****

On the Texture Bias for Few-Shot CNN Segmentation

Reza Azad, Abdur R. Fayjie, Claude Kauffmann, Ismail Ben Ayed, Marco Pedersoli, Jose Dolz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2674-2683

Despite the initial belief that Convolutional Neural Networks (CNNs) are driven by shapes to perform visual recognition tasks, recent evidence suggests that texture bias in CNNs provides higher performing and more robust models. This contrasts with the perceptual bias in the human visual cortex, which has a stronger preference towards shape components. Perceptual differences may explain why CNNs achieve human-level performance when large labeled datasets are available, but their performance significantly degrades in low-labeled data scenarios, such as few-shot semantic segmentation. To remove the texture bias in the context of few-shot learning, we propose a novel architecture that integrates a set of Difference of Gaussians (DoG) to attenuate high-frequency local components in the feature space. This produces a set of modified feature maps, whose high-frequency components are diminished at different standard deviation values of the Gaussian distribution in the spatial domain. As this results in multiple feature maps for a single image, we employ a bi-directional convolutional long-short-term-memory to efficiently merge the multi scale-space representations. We perform extensive experiments on three well-known few-shot segmentation benchmarks --Pascal i5, COCO-20i and FSS-1000-- and demonstrate that our method outperforms state-of-the-art approaches in two datasets under the same conditions.

****

Revisiting Street-to-Aerial View Image Geo-Localization and Orientation Estimati on

Sijie Zhu, Taojiannan Yang, Chen Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 756-765

Street-to-aerial image geo-localization, which matches a query street-view image to the GPS-tagged aerial images in a reference set, has attracted increasing attention recently. In this paper, we revisit this problem and point out the ignored issue about image alignment information. We show that the performance of a simple Siamese network is highly dependent on the alignment setting and the comparison of previous works can be unfair if they have different assumptions. Instead of focusing on the feature extraction under the alignment assumption, we show that improvements in metric learning techniques significantly boost the performance regardless of the alignment. Without leveraging the alignment information, our pipeline outperforms previous works on both panorama and cropped datasets. Furthermore, we conduct visualization to help understand the learned model and the effect of alignment information. With our discovery on the approximate rotation-invariant activation map, we propose a novel orientation estimation method that achieves state-of-the-art results on the CVUSA dataset.

****

Appending Adversarial Frames for Universal Video Attack

Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, Qi Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3199-3208

This paper investigates the problem of generating adversarial examples for video classification. We project all videos onto a semantic space and a perception space, and point out that adversarial attack is to find a counterpart which is close to the target in the perception space but far from the target in the semantic space. Based on this formulation, we notice that conventional attacking methods mostly used Euclidean distance to measure the perception space, but we propose to make full use of the property of videos and assume a modified video with a few consecutive frames replaced by dummy contents (e.g., a black frame with texts of `thank you for watching' on it) to be close to the original video in the perception space though they have a large Euclidean gap. This leads to a new attack approach which only adds perturbations on the newly-added frames. We show its high success rates in attacking six state-of-the-art video classification networks, as well as its universality, i.e., transferring well across videos and models.

****

WDNet: Watermark-Decomposition Network for Visible Watermark Removal

Yang Liu, Zhen Zhu, Xiang Bai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3685-3693

Visible watermarks are widely-used in images to protect copyright ownership. Analyzing watermark removal helps to reinforce the anti-attack techniques in an adversarial way. Current removal methods normally leverage image-to-image translation techniques. Nevertheless, the uncertainty of the size, shape, color and transparency of the watermarks set a huge barrier for these methods. To combat this, we combine traditional watermarked image decomposition into a two-stage generator, called Watermark-Decomposition Network (WDNet), where the first stage predicts a rough decomposition from the whole watermarked image and the second stage specifically centers on the watermarked area to refine the removal results. The decomposition formulation enables WDNet to separate watermarks from the images rather than simply removing them. We further show that these separated watermarks can serve as extra nutrients for building a larger training dataset and further improving removal performance. Besides, we construct a large-scale dataset named CLWD, which mainly contains colored watermarks, to fill the vacuum of colored watermark removal dataset. Extensive experiments on the public gray-scale dataset LVW and CLWD consistently show that the proposed WDNet outperforms the state-of-the-art approaches both in accuracy and efficiency.

****

High-Quality Frame Interpolation via Tridirectional Inference

Jinsoo Choi, Jaesik Park, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 596-604

Videos have recently become an omnipresent form of media, gathering much attention from industry as well as academia. In the video enhancement field, video frame interpolation is a long-studied topic that has dramatically improved due to the advancement of deep convolutional neural networks (CNN). However, conventional approaches utilizing two successive frames often exhibit ghosting or tearing artifacts for moving objects. We argue that this phenomenon comes from the lack of reliable information provided only by two frames. With this motivation, we propose a frame interpolation method by utilizing tridirectional information obtained from three input frames. Information extracted from triplet frames allows our model to learn rich and reliable inter-frame motion representations, including subtle nonlinear movement, which can be easily trained via any video frames in a self-supervised manner. We demonstrate that our method generalizes well to high-resolution content by evaluating on FHD resolution, and illustrates our approach's effectiveness via comparison to state-of-the-art methods on challenging video content.

****

DualSANet: Dual Spatial Attention Network for Iris Recognition

Kai Yang, Zihao Xu, Jingjing Fei; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 889-897

Compared with other human biosignatures, iris has more advantages on accuracy, invariability and robustness. However, the performance of existing common iris recognition algorithms is still far from expectations of the community. Although some researchers have attempted to utilize deep learning methods which are superior to traditional methods, it is worth exploring better CNN network architecture. In this paper, we propose a novel network architecture based on the dual spatial attention mechanism for iris recognition, called DualSANet. Specifically, the proposed architecture can generate multi-level spatially corresponding feature representations via an encoder-decoder structure. In the meantime, we also propose a new spatial attention feature fusion module, so as to ensemble these features more effectively. Based on these, our architecture can generate dual feature representations which have complementary discriminative information. Extensive experiments are conducted on CASIA-IrisV4-Thousand, CASIA-IrisV4-Distance, and IITD datasets. The experimental results show that our method achieves superior performance compared with the state-of-the-arts.

****

Joint Visual-Temporal Embedding for Unsupervised Learning of Actions in Untrimmed Sequences

Rosaura G. VidalMata, Walter J. Scheirer, Anna Kukleva, David Cox, Hilde Kuehne; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1238-1247

Understanding the structure of complex activities in untrimmed videos is a challenging task in the area of action recognition. One problem here is that this task usually requires a large amount of hand-annotated minute- or even hour-long video data, but annotating such data is very time consuming and can not easily be automated or scaled. To address this problem, this paper proposes an approach for the unsupervised learning of actions in untrimmed video sequences based on a joint visual-temporal embedding space. To this end, we combine a visual embedding based on a predictive U-Net architecture with a temporal continuous function. The resulting representation space allows detecting relevant action clusters based on their visual as well as their temporal appearance. The proposed method is evaluated on three standard benchmark datasets, Breakfast Actions, INRIA YouTube Instructional Videos, and 50 Salads. We show that the proposed approach is able to provide a meaningful visual and temporal embedding out of the visual cues present in contiguous video frames and is suitable for the task of unsupervised temporal segmentation of actions.

****

Efficient Video Annotation With Visual Interpolation and Frame Selection Guidance

Alina Kuznetsova, Aakrati Talati, Yiwen Luo, Keith Simmons, Vittorio Ferrari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (

WACV), 2021, pp. 3070-3079

We introduce a unified framework for generic video annotation with bounding boxes. Video annotation is a longstanding problem, as it is a tedious and time-consuming process. We tackle two important challenges of video annotation: (1) automatic temporal interpolation and extrapolation of bounding boxes provided by a human annotator on a subset of all frames, and (2) automatic selection of frames to annotate manually. Our contribution is two-fold: first, we propose a model that has both interpolating and extrapolating capabilities; second, we propose a guiding mechanism that sequentially generates suggestions for what frame to annotate next, based on the annotations made previously. We extensively evaluate our approach on several challenging datasets in simulation and demonstrate a reduction in terms of the number of manual bounding boxes drawn by 60% over linear interpolation and by 35% over an off-the shelf tracker. Moreover, we also show 10% annotation time improvement over a state-of-the-art method for video annotation with bounding boxes [25]. Finally, we run human annotation experiments and provide extensive analysis of the results, showing that our approach reduces actual measured annotation time by 50% compared to commonly used linear interpolation.

****

DocVQA: A Dataset for VQA on Document Images

Minesh Mathew, Dimosthenis Karatzas, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2200-2209

We present a new dataset for Visual Question Answering (VQA) on document images called DocVQA. The dataset consists of 50,000 questions defined on 12,000+ document images. Detailed analysis of the dataset in comparison with similar datasets for VQA and reading comprehension is presented. We report several baseline results by adopting existing VQA and reading comprehension models. Although the existing models perform reasonably well on certain types of questions, there is large performance gap compared to human performance (94.36% accuracy). The models need to improve specifically on questions where understanding structure of the document is crucial. The dataset, code and leaderboard are available at docvqa.org

****

FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation

Kimmo Karkkainen, Jungseock Joo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1548-1558

Existing public face image datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented. The models trained from such datasets suffer from inconsistent classification accuracy, which limits the applicability of face analytic systems to non-White race groups. To mitigate the race bias problem in these datasets, we constructed a novel face image dataset containing 108,501 images which is balanced on race. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Images were collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. Evaluations were performed on existing face attribute datasets as well as novel image datasets to measure the generalization performance. We find that the model trained from our dataset is substantially more accurate on novel datasets and the accuracy is consistent across race and gender groups. We also compare several commercial computer vision APIs and report their balanced accuracy across gender, race, and age groups.

****

CPM R-CNN: Calibrating Point-Guided Misalignment in Object Detection

Bin Zhu, Qing Song, Lu Yang, Zhihui Wang, Chun Liu, Mengjie Hu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3248-3257

In object detection, offset-guided and point-guided regression dominate anchor-based and anchor-free method separately. Recently, point-guided approach is introduced to anchor-based method. However, we observe points predicted by this way are misaligned with matched region of proposals and score of localization, causing a notable gap in performance. In this paper, we propose CPM R-CNN which contains three efficient modules to optimize anchor-based point-guided method. Accordi ng to sufficient evaluations on the COCO dataset, CPM R-CNN is demonstrated efficient to improve the localization accuracy by calibrating mentioned misalignment. Compared with Faster R-CNN and Grid R-CNN based on ResNet-101 with FPN, our approach can substantially improve detection mAP by 3.3% and 1.5% respectively without whistles and bells. Moreover, our best model achieves improvement by a large margin to 49.9% on COCO test-dev. Code and models will be publicly available.

****

Revisiting Adaptive Convolutions for Video Frame Interpolation

Simon Niklaus, Long Mai, Oliver Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1099-1109

Video frame interpolation, the synthesis of novel views in time, is an increasingly popular research direction with many new papers further advancing the state of the art. But as each new method comes with a host of variables that affect the interpolation quality, it can be hard to tell what is actually important for this task. In this work, we show, somewhat surprisingly, that it is possible to achieve near state-of-the-art results with an older, simpler approach, namely adaptive separable convolutions, by a subtle set of low level improvements. In doing so, we propose a number of intuitive but effective techniques to improve the frame interpolation quality, which also have the potential to other related applications of adaptive convolutions such as burst image denoising, joint image filtering, or video prediction.

****

Do We Really Need Gold Samples for Sample Weighting Under Label Noise?

Aritra Ghosh, Andrew Lan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3922-3931

Learning with labels noise has gained significant traction recently due to the sensitivity of deep neural networks under label noise under common loss functions. Losses that are theoretically robust to label noise, however, often makes training difficult. Consequently, several recently proposed methods, such as Meta-Weight-Net (MW-Net), use a small number of unbiased, clean samples to learn a weighting function that downweights samples that are likely to have corrupted labels under the meta-learning framework. However, obtaining such a set of clean samples is not always feasible in practice. In this paper, we analytically show that one can easily train MW-Net without access to clean samples simply by using a loss function that is robust to label noise, such as mean absolute error, as the meta objective to train the weighting network. We experimentally show that our method beats all existing methods that do not use clean samples and performs on-par with methods that use gold samples on benchmark datasets across various noise types and noise rates.

****

Part Segmentation of Unseen Objects Using Keypoint Guidance

Shujon Naha, Qingyang Xiao, Prianka Banik, Md. Alimoor Reza, David J. Crandall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1742-1750

While object part segmentation is useful for many applications, typical approaches require a large amount of labeled data to train a model for good performance. To reduce the labeling effort, weak supervision cues such as object keypoints have been used to generate pseudo-part annotations which can subsequently be used to train larger models. However, previous weakly-supervised part segmentation methods require the same object classes during both training and testing. We propose a new model to use key-point guidance for segmenting parts of novel object classes given that they have similar structures as seen objects --different types of four-legged animals, for example. We show that a non-parametric template matching approach is more effective than pixel classification for part segmentation, especially for small or less frequent parts. To evaluate the generalizability of our approach, we introduce two new datasets that contain 200 quadrupeds in total with both key-point and part segmentation annotations. We show that our approach can outperform existing models by a large mar-gin on the novel object part segmentation task using limited part segmentation labels during training.

****

Intra-Class Part Swapping for Fine-Grained Image Classification

Lianbo Zhang, Shaoli Huang, Wei Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3209-3218

Recent works such as Mixup and Cutmix have demonstrated the effectiveness of augmenting training data for deep models. These methods generate new data by generally blending random image contents and mixing their labels proportionally. However, this strategy tends to produce unreasonable training samples for fine-grained recognition, leading to limited improvement. This is because mixing random image contents may potentially produce images containing destructed object structures. Further, as the category differences mainly reside in small part regions, mixing labels proportionally to the number of mixed pixels might result in label noisy problem. To augment more reasonable training data, we propose Intra-class Part Swapping (InPS) that produces new data by performing attention-guided content swapping on input pairs from the same class. Compared with previous approaches, InPS avoids introducing noisy labels and ensures a likely holistic structure of objects in generated images. We demonstrate InPS outperforms the most recent augmentation approaches in both fine-grained recognition and weakly object localization. Further, by simply incorporating the mid-level feature learning, our proposed method achieves state-of-the-art performance in the literature while maintaining the simplicity and inference efficiency. Our code is publicly available.

****

Rotate to Attend: Convolutional Triplet Attention Module

Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, Qibin Hou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3139-3148

Benefiting from the capability of building inter-dependencies among channels or spatial locations, attention mechanisms have been extensively studied and broadly used in a variety of computer vision tasks recently. In this paper, we investigate light-weight but effective attention mechanisms and present triplet attention, a novel method for computing attention weights by capturing cross-dimension interaction using a three-branch structure. For an input tensor, triplet attention builds inter-dimensional dependencies by the rotation operation followed by residual transformations and encodes inter-channel and spatial information with negligible computational overhead. Our method is simple as well as efficient and can be easily plugged into classic backbone networks as an add-on module. We demonstrate the effectiveness of our method on various challenging tasks including image classification on ImageNet-1k and object detection on MSCOCO and PASCAL VOC datasets. Furthermore, we provide extensive in-sight into the performance of triplet attention by visually inspecting the GradCAM and GradCAM++ results. The empirical evaluation of our method supports our intuition on the importance of capturing dependencies across dimensions when computing attention weights. Code for this paper can be publicly accessed at https://github.com/LandskapeAI/triplet-attention.

****

A Deflation Based Fast and Robust Preconditioner for Bundle Adjustment

Shrutimoy Das, Siddhant Katyan, Pawan Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1782-1789

The bundle adjustment(BA) problem is formulated as a non linear least squares problem, which requires the solution of a linear system. For solving this system, we present the design and implementation of a fast preconditioned solver. The proposed preconditioner is based on the deflation of the largest eigenvalues of the Hessian. We also derive an estimate of the condition number of the preconditioned system. Numerical experiments on problems from the BAL dataset suggest that our solver is the fastest, sometimes, by a factor of five, when compared to the current state-of-the-art solvers for bundle adjustment.

****

Integrating Human Gaze Into Attention for Egocentric Activity Recognition

Kyle Min, Jason J. Corso; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1069-1078

It is well known that human gaze carries significant information about visual at tention. However, there are three main difficulties in incorporating the gaze data in an attention mechanism of deep neural networks: 1) the gaze fixation points are likely to have measurement errors due to blinking and rapid eye movements; 2) it is unclear when and how much the gaze data is correlated with visual attention; and 3) gaze data is not always available in many real-world situations. In this work, we introduce an effective probabilistic approach to integrate human gaze into spatiotemporal attention for egocentric activity recognition. Specifically, we represent the locations of gaze fixation points as structured discrete latent variables to model their uncertainties. In addition, we model the distribution of gaze fixations using a variational method. The gaze distribution is learned during the training process so that the ground-truth annotations of gaze locations are no longer needed in testing situations since they are predicted from the learned gaze distribution. The predicted gaze locations are used to provide informative attentional cues to improve the recognition performance. Our method outperforms all the previous state-of-the-art approaches on EGTEA, which is a large-scale dataset for egocentric activity recognition provided with gaze measurements. We also perform an ablation study and qualitative analysis to demonstrate that our attention mechanism is effective.

****

StressNet: Detecting Stress in Thermal Videos

Satish Kumar, A S M Iftekhar, Michael Goebel, Tom Bullock, Mary H. MacLean, Michael B. Miller, Tyler Santander, Barry Giesbrecht, Scott T. Grafton, B.S. Manjunath; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 999-1009

Precise measurement of physiological signals is critical for the effective monitoring of human vital signs. Recent developments in computer vision have demonstrated that signals such as pulse rate and respiration rate can be extracted from digital video of humans, increasing the possibility of contact-less monitoring. This paper presents a novel approach to obtaining physiological signals and classifying stress states from thermal video. The proposed net-work "StressNet", features a hybrid emission representation model that models the direct emission and absorption of heat by the skin and underlying blood vessels. This results in an information-rich feature representation of the face, which is used by spatio-temporal networks for recon-structing the ISTI (Initial Systolic Time Interval : a measure of change in cardiac sympathetic activity that is considered to be a quantitative index of stress in humans). The recon-structed ISTI signal is fed to a stress-detection model to detect and classify the individual's stress state (i.e. stress or no stress). A detailed evaluation demonstrates that Stress-Net achieves a mean square error of 5.845 ms for predicting the ISTI signal and an average precision of 0.842 for stress detection.

****

Driver Anomaly Detection: A Dataset and Contrastive Learning Approach

Okan Kopuklu, Jiapeng Zheng, Hang Xu, Gerhard Rigoll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 91-100

Distracted drivers are more likely to fail to anticipate hazards, which result in car accidents. Therefore, detecting anomalies in drivers' actions (i.e., any action deviating from normal driving) contains the utmost importance to reduce driver-related accidents. However, there are unbounded many anomalous actions that a driver can do while driving, which leads to an `open set recognition' problem. Accordingly, instead of recognizing a set of anomalous actions that are commonly defined by previous dataset providers, in this work, we propose a contrastive learning approach to learn a metric to differentiate normal driving from anomalous driving. For this task, we introduce a new video-based benchmark, the Driver Anomaly Detection (DAD) dataset, which contains normal driving videos together with a set of anomalous actions in its training set. In the test set of the DAD dataset, there are unseen anomalous actions that still need to be winnowed out from normal driving. Our method reaches 0.9673 AUC on the test set, demonstrating the effectiveness of the contrastive learning approach on the anomaly detection task. Our dataset, codes and pre-trained models are publicly available.

****

TResNet: High Performance GPU-Dedicated Architecture
Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, Itamar Friedman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1400-1409

Many deep learning models, developed in recent years, reach higher ImageNet accuracy than ResNet50, with fewer or comparable FLOPs count. While FLOPs are often seen as a proxy for network efficiency, when measuring actual GPU training and inference throughput, vanilla ResNet50 is usually significantly faster than its recent competitors, offering better throughput-accuracy trade-off. In this work, we introduce a series of architecture modifications that aim to boost neural networks' accuracy, while retaining their GPU training and inference efficiency. We first demonstrate and discuss the bottlenecks induced by FLOPs oriented optimizations. We then suggest alternative designs that better utilize GPU structure and assets. Finally, we introduce a new family of GPU-dedicated models, called TResNet, which achieves better accuracy and efficiency than previous ConvNets. Using a TResNet model, with similar GPU throughput to ResNet50, we reach 80.8% top-1 accuracy on ImageNet. Our TResNet models also transfer well and achieve state-of-the-art accuracy on competitive single-label classification datasets such as Stanford Cars (96.0%), CIFAR-10 (99.0%), CIFAR-100 (91.5%) and Oxford-Flowers (99.1%). TResNet models also achieve state-of-the-art results on a multi-label classification task, and perform well on object detection.

****

The MECCANO Dataset: Understanding Human-Object Interactions From Egocentric Videos in an Industrial-Like Domain
Francesco Ragusa, Antonino Furnari, Salvatore Livatino, Giovanni Maria Farinella; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1569-1578

Wearable cameras allow to collect images and videos of humans interacting with the world. While human-object interactions have been thoroughly investigated in third person vision, the problem has been understudied in egocentric settings and in industrial scenarios. To fill this gap, we introduce MECCANO, the first dataset of egocentric videos to study human-object interactions in industrial-like settings. MECCANO has been acquired by 20 participants who were asked to build a motorbike model, for which they had to interact with tiny objects and tools. The dataset has been explicitly labeled for the task of recognizing human-object interactions from an egocentric perspective. Specifically, each interaction has been labeled both temporally (with action segments) and spatially (with active object bounding boxes). With the proposed dataset, we investigate four different tasks including 1) action recognition, 2) active object detection, 3) active object recognition and 4) egocentric human-object interaction detection, which is a revisited version of the standard human-object interaction detection task. Baseline results show that the MECCANO dataset is a challenging benchmark to study egocentric human-object interactions in industrial-like scenarios. We publicy release the dataset at https://iplab.dmi.unict.it/MECCANO/.

****

Towards Enhancing Fine-Grained Details for Image Matting
Chang Liu, Henghui Ding, Xudong Jiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 385-393

In recent years, deep natural image matting has been rapidly evolved by extracting high-level contextual features into the model. However, most current methods still have difficulties with handling tiny details, like hairs or furs. In this paper, we argue that recovering these microscopic details relies on low-level but high-definition texture features. However, these features are downsampled in a very early stage in current encoder-decoder-based models, resulting in the loss of microscopic details . To address this issue, we design a deep image matting model to enhance fine-grained details. Our model consists of two parallel paths: a conventional encoder-decoder Semantic Path and an independent downsampling-free Textural Compensate Path (TCP). The TCP is proposed to extract fine-grained details such as lines and edges in the original image size, which greatly enhances the fineness of prediction. Meanwhile, to leverage the benefits of high-lev el context, we propose a feature fusion unit(FFU) to fuse multi-scale features from the semantic path and inject them into the TCP. In addition, we have observed that poorly annotated trimaps severely affect the performance of the model. Thus we further propose a novel term in loss function and a trimap generation method to improve our model's robustness to the trimaps. The experiments show that our method outperforms previous start-of-the-art methods on the Composition-1k dataset.

****

SHAD3S: A Model to Sketch, Shade and Shadow

Raghav Brahmadesam Venkataramaiyer, Abhishek Joshi, Saisha Narang, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3616-3625

Hatching is a common method used by artists to accentuate the third dimension of a sketch, and to illuminate the scene. Our system SHAD3S attempts to compete with a human at hatching generic three-dimensional (3D) shapes, and also tries to assist her in a form exploration exercise. The novelty of our approach lies in the fact that we make no assumptions about the input other than that it represents a 3D shape, and yet, given a contextual information of illumination and texture, we synthesise an accurate hatch pattern over the sketch, without access to 3d or pseudo 3D. In the process, we contribute towards a) a cheap yet effective method to synthesise a sufficiently large high fidelity dataset, pertinent to task; b) creating a pipeline with conditional generative adversarial network (CGAN); and c) creating an interactive utility with GIMP, that is a tool for artists to engage with automated hatching or a form-exploration exercise. User evaluation of the tool suggests that the model performance does generalise satisfactorily over diverse input, both in terms of style as well as shape. A simple comparison of inception scores suggest that the generated distribution is as diverse as the ground truth.k

****

Learning Shape Representations for Person Re-Identification Under Clothing Change

Yu-Jhe Li, Xinshuo Weng, Kris M. Kitani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2432-2441

Person re-identification (re-ID) aims to recognize instances of the same person contained in multiple images taken across different cameras. Existing methods for re-ID tend to rely heavily on the assumption that both query and gallery images of the same person have the same clothing. Unfortunately, this assumption may not hold for datasets captured over long periods of time. To tackle the re-ID problem in the context of clothing changes, we propose a novel representation learning method which is able to generate a shape-based feature representation that is invariant to clothing. We call our model the Clothing Agnostic Shape Extraction Network (CASE-Net). CASE-Net learns a representation of a person that depends primarily on shape via adversarial learning and feature disentanglement. Quantitative and qualitative results across 5 datasets (Div-Market, Market1501, three large-scale datesets under clothing changes) show our approach makes significant improvements over prior state-of-the-art approaches.

****

Improved Techniques for Training Single-Image GANs

Tobias Hinz, Matthew Fisher, Oliver Wang, Stefan Wermter; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1300-1309

Recently there has been an interest in the potential of learning generative models from a single image, as opposed to from a large dataset. This task is of significance, as it means that generative models can be used in domains where collecting a large dataset is not feasible. However, training a model capable of generating realistic images from only a single sample is a difficult problem. In this work, we conduct a number of experiments to understand the challenges of training these methods and propose some best practices that we found allowed us to generate improved results over previous work. One key piece is that, unlike prior single image generation methods, we concurrently train several stages in a sequen tial multi-stage manner, allowing us to learn models with fewer stages of increasing image resolution. Compared to a recent state of the art baseline, our model is up to six times faster to train, has fewer parameters, and can better capture the global structure of images.

****

Visual Tracking of Deepwater Animals Using Machine Learning-Controlled Robotic Underwater Vehicles

Kakani Katija, Paul L. D. Roberts, Joost Daniels, Alexandra Lapides, Kevin Barnard, Mike Risi, Ben Y. Ranaan, Benjamin G. Woodward, Jonathan Takahashi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 860-869

The ocean is a vast three-dimensional space that is poorly explored and understood, and harbors unobserved life and processes that are vital to ecosystem function. To fully interrogate the space, novel algorithms and robotic platforms are required to scale up observations. Locating animals of interest and extended visual observations in the water column are particularly challenging objectives. Towards that end, we present a novel Machine Learning-integrated Tracking (or ML-Tracking) algorithm for underwater vehicle control that builds on the class of algorithms known as tracking-by-detection. By coupling a multi-object detector (trained on in situ underwater image data), a 3D stereo tracker, and a supervisor module to oversee the mission, we show how ML-Tracking can create robust tracks needed for long duration observations, as well as enable fully automated acquisition of objects for targeted sampling. Using a remotely operated vehicle as a proxy for an autonomous underwater vehicle, we demonstrate continuous input from the ML-Tracking algorithm to the vehicle controller during a record, 5+ hr continuous observation of a midwater gelatinous animal known as a siphonophore. These efforts clearly demonstrate the potential that tracking-by-detection algorithms can have on exploration in unexplored environments and discovery of undiscovered life in our ocean.

****

Few-Shot Font Style Transfer Between Different Languages

Chenhao Li, Yuta Taniguchi, Min Lu, Shin'ichi Konomi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 433-442

In this paper, we propose a novel model FTransGAN that can transfer font styles between different languages by observing only a few samples. The automatic generation of a new font library is a challenging task and has been attracting many researchers' interests. Most previous works addressed this problem by transferring the style of the given subset to the content of unseen ones. Nevertheless, they only focused on the font style transfer in the same language. In many tasks, we need to learn the font information from one language and then apply it to other languages. It's difficult for the existing methods to do such tasks. To solve this problem, we specifically design our network into a multi-level attention form to capture both local and global features of the style images. To verify the generative ability of our model, we construct an experimental font dataset which includes 847 fonts, each of them containing English and Chinese characters with the same style. Experimental results show that compared with the state-of-the-art models, our model generates 80.3% of all user preferred images.

****

A Learning-Based Approach to Parametric Rotoscoping of Multi-Shape Systems

Luis Bermudez, Nadine Dabby, Yingxi Adelle Lin, Sara Hilmarsdottir, Narayan Sundararajan, Swarnendu Kar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 776-785

Rotoscoping of facial features is often an integral part of Visual Effects post-production, where the parametric contours created by artists need to be highly detailed, consist of multiple interacting components, and involve significant manual supervision. Yet those assets are usually discarded after compositing and hardly reused. In this paper, we present the first methodology to learn from these assets. With only a few manually rotoscoped shots, we identify and extract semantically consistent and task specific landmark points and re-vectorize the roto shapes based on these landmarks. We then train two separate models -- one to pre

dict landmarks based on a rough crop of the face region, and the other to predict the roto shapes using only the inferred landmarks from the first model. In preliminary production testing, 26% of shots rotoscoped using our tool were able to be used with no adjustment, and another 47% were able to be used with minor adjustments. This represents a significant time savings for the studio, as artists are able to rotoscope almost 73% of their shots with no manual rotoscoping and some spline adjustment. This paper presents a novel application of machine learning to professional interactive rotoscoping, a methodology to convert unstructured roto shapes into a self-annotated, trainable dataset that can be harnessed to make accurate predictions on future shots of a similar object, and a limited dataset of rotoscoped multi-shape fine feature systems from a real film production.
****

Task-Assisted Domain Adaptation With Anchor Tasks
Zhizhong Li, Linjie Luo, Sergey Tulyakov, Qieyun Dai, Derek Hoiem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2989-2998
Some tasks, such as surface normals or single-view depth estimation, require per-pixel ground truth that is difficult to obtain on real images but easy to obtain on synthetic. However, models learned on synthetic images often do not generalize well to real images due to the domain shift. Our key idea to improve domain adaptation is to introduce a separate anchor task (such as facial landmarks) whose annotations can be obtained at no cost or are already available on both synthetic and real datasets. To further leverage the implicit relationship between the anchor and main tasks, we apply our HeadFreeze technique that learns the cross-task guidance on the source domain with the final network layers, and use it on the target domain. We evaluate our methods on surface normal estimation on two pairs of datasets (indoor scenes and faces) with two kinds of anchor tasks (semantic segmentation and facial landmarks). We show that blindly applying domain adaptation or training the auxiliary task on only one domain may hurt performance, while using anchor tasks on both domains is better behaved. Our HeadFreeze technique outperforms competing approaches, reaching performance in facial images on par with a recently popular surface normal estimation method using shape from shading domain knowledge.
****

Towards Fair Cross-Domain Adaptation via Generative Learning
Tongxin Wang, Zhengming Ding, Wei Shao, Haixu Tang, Kun Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 454-463
Domain Adaptation (DA) targets at adapting a model trained over the well-labeled source domain to the unlabeled target domain lying in different distributions. Existing DA normally assumes the well-labeled source domain is class-wise balanced, which means the size per source class is relatively similar. However, in real-world applications, labeled samples for some categories in the source domain could be extremely few due to the difficulty of data collection and annotation, which leads to decreasing performance over target domain on those few-shot categories. To perform fair cross-domain adaptation and boost the performance on these minority categories, we develop a novel Generative Few-shot Cross-domain Adaptation (GFCA) algorithm for fair cross-domain classification. Specifically, generative feature augmentation is explored to synthesize effective training data for few-shot source classes, while effective cross-domain alignment aims to adapt knowledge from source to facilitate the target learning. Experimental results on two large cross-domain visual datasets demonstrate the effectiveness of our proposed method on improving both few-shot and overall classification accuracy comparing with the state-of-the-art DA approaches.
****

Neuron Matching in C. elegans With Robust Approximate Linear Regression Without Correspondence
Amin Nejatbakhsh, Erdem Varol; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2837-2846
We propose methods for estimating correspondence between two point sets under th

e presence of outliers in both the source and target sets. The proposed algorithms expand upon the theory of the regression without correspondence problem to estimate transformation coefficients using unordered multisets of covariates and responses. Previous theoretical analysis of the problem has been done in a setting where the responses are a complete permutation of the regressed covariates. This paper expands the problem setting by analyzing the cases where only a subset of the responses is a permutation of the regressed covariates in addition to some covariates possibly being adversarial outliers. We term this problem robust regression without correspondence and provide several algorithms based on random sample consensus for exact and approximate recovery in a noiseless and noisy one-dimensional setting as well as an approximation algorithm for multiple dimensions. The theoretical guarantees of the algorithms are verified in simulated data. We demonstrate an important computational neuroscience application of the proposed framework by demonstrating its effectiveness in a Caenorhabditis elegans neuron matching problem where the presence of outliers in both the source and target nematodes is a natural tendency.

****

H2O-Net: Self-Supervised Flood Segmentation via Adversarial Domain Adaptation and Label Refinement
Peri Akiva, Matthew Purri, Kristin Dana, Beth Tellman, Tyler Anderson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 111-122

Accurate flood detection in near real time via high resolution, high latency satellite imagery is essential to prevent loss of lives by providing quick and actionable information. Instruments and sensors useful for flood detection are only available in low resolution, low latency satellites with region re-visit periods of up to 16 days, making flood alerting systems that use such satellites unreliable. This work presents H2O-Network, a self supervised deep learning method to segment floods from satellites and aerial imagery by bridging domain gap between low and high latency satellite and coarse-to-fine label refinement. H2O-Net learns to synthesize signals highly correlative with water presence as a domain adaptation step for semantic segmentation in high resolution satellite imagery. Our work also proposes a self-supervision mechanism, which does not require any hand annotation, used during training to generate high quality ground truth data. We demonstrate that H2O-Net outperforms the state-of-the-art semantic segmentation methods on satellite imagery by 10% and 12% pixel accuracy and mIoU respectively for the task of flood segmentation. We emphasize the generalizability of our model by transferring model weights trained on satellite imagery to drone imagery, a highly different sensor and domain.

****

EvidentialMix: Learning With Combined Open-Set and Closed-Set Noisy Labels
Ragav Sachdeva, Filipe R. Cordeiro, Vasileios Belagiannis, Ian Reid, Gustavo Carneiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3607-3615

The efficacy of deep learning depends on large-scale data sets that have been carefully curated with reliable data acquisition and annotation processes. However, acquiring such large-scale data sets with precise annotations is very expensive and time-consuming, and the cheap alternatives often yield data sets that have noisy labels. The field has addressed this problem by focusing on training models under two types of label noise: 1) closed-set noise, where some training samples are incorrectly annotated to a training label other than their known true class; and 2) open-set noise, where the training set includes samples that possess a true class that is (strictly) not contained in the set of known training labels. In this work, we study a new variant of the noisy label problem that combines the open-set and closed-set noisy labels, and introduce a benchmark evaluation to assess the performance of training algorithms under this setup. We argue that such problem is more general and better reflects the noisy label scenarios in practice. Furthermore, we propose a novel algorithm, called EvidentialMix, that addresses this problem and compare its performance with the state-of-the-art methods for both closed-set and open-set noise on the proposed benchmark. Our resul ts show that our method produces superior classification results and better feature representations than previous state-of-the-art methods. The code is available at https://github.com/ragavsachdeva/EvidentialMix.

****

Facial Emotion Recognition With Noisy Multi-Task Annotations

Siwei Zhang, Zhiwu Huang, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 21-31

Human emotions can be inferred from facial expressions. However, the annotations of facial expressions are often highly noisy in common emotion coding models, including categorical and dimensional ones. To reduce human labelling effort on multi-task labels, we introduce a new problem of facial emotion recognition with noisy multi-task annotations. For this new problem, we suggest a formulation from the point of joint distribution match view, which aims at learning more reliable correlations among raw facial images and multi-task labels, resulting in the reduction of noise influence. In our formulation, we exploit a new method to enable the emotion prediction and the joint distribution learning in a unified adversarial learning game. Evaluation throughout extensive experiments studies the real setups of the suggested new problem, as well as the clear superiority of the proposed method over the state-of-the-art competing methods on either the synthetic noisy labeled CIFAR-10 or practical noisy multi-task labeled RAF and AffectNet. The code is available at https://github.com/sanweiliti/noisyFER.

****

Fast Kernelized Correlation Filter Without Boundary Effect

Ming Tang, Linyu Zheng, Bin Yu, Jinqiao Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2999-3008

In recent years, correlation filter based trackers (CF trackers) have attracted much attention from the vision community because of their top performance in both localization accuracy and efficiency. The society of visual tracking, however, still needs to deal with the following difficulty on CF trackers: avoiding or eliminating the boundary effect completely, in the meantime, exploiting non-linear kernels and running efficiently. In this paper, we propose a fast kernelized correlation filter without boundary effect (nBEKCF) to solve this problem. To avoid the boundary effect thoroughly, a set of real and dense patches is sampled through the traditional sliding window and used as the training samples to train nBEKCF to fit a Gaussian response map. Non-linear kernels can be applied naturally in nBEKCF due to its different theoretical foundation from the existing CF trackers'. To achieve the fast training and detection, a set of cyclic bases is introduced to construct the filter. Two algorithms, ACSII and CCIM, are developed to significantly accelerate the calculation of kernel correlation matrices. ACSII and CCIM fully exploit the density of training samples and cyclic structure of bases, and totally run in space domain. The efficiency of CCIM exceeds that of the FFT counterpart remarkably in our task. Extensive experiments on six public datasets, OTB-2013, OTB-2015, NfS, VOT2018, GOT10k, and TrackingNet, show that compared to the CF trackers designed to relax the boundary effect, BACF and SRDCF, our nBEKCF achieves higher localization accuracy without tricks, in the meanwhile, runs at higher FPS.

****

PDAN: Pyramid Dilated Attention Network for Action Detection

Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2970-2979

Handling long and complex temporal information is an important factor for action detection tasks. This challenge is further aggravated by densely distributed actions in untrimmed videos. Previous action detection methods are failing in selecting the key temporal information in videos of long length. To this end, we introduce the Dilated Attention Layer (DAL). Compared to previous temporal convolution layer, DAL allocates attentional weights to each feature in the kernel, which enables DAL to learn better local representation across time. Furthermore, DAL when accompanied by dilated kernels is able to learn a global representation of

several minutes long videos which is crucial for the task of action detection. Finally, we introduce Pyramid Dilated Attention Network (PDAN) which is build up on DAL. With the help of DAL combining with dilation and residual links, PDAN can model short-term and long-term temporal relations simultaneously by focusing on local segments at the level of low and high temporal receptive fields. This property enables PDAN to handle complex temporal relations between different action instances in long untrimmed videos. To corroborate the effectiveness and robustness of our proposed method, we evaluate it on three densely annotated, multi-label datasets: MultiTHUMOS, Charades and an Inhouse dataset, outperforming the state-of-the-art results.
****

ChartOCR: Data Extraction From Charts Images via a Deep Hybrid Framework
Junyu Luo, Zekun Li, Jinpeng Wang, Chin-Yew Lin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1917-1925
Chart images are commonly used for data visualization. Automatically reading the chart values is a key step for chart content understanding. Charts have a lot of variations in style (e.g. bar chart, line chart, pie chart and etc.), which makes pure rule-based data extraction methods difficult to handle. However, it is also improper to directly apply end-to-end deep learning solutions since these methods usually deal with specific types of charts. In this paper, we propose an unified method ChartOCR to extract data from various types of charts. We show that by combing deep framework and rule-based methods, we can achieve a satisfying generalization ability and obtain accurate and semantic-rich intermediate results. Our method extracts the key points that define the chart components. By adjusting the prior rules, the framework can be applied to different chart types. Experiments show that our method achieves state-of-the-art performance with fast processing speed on two public datasets. Besides, we also introduce and evaluate on a large dataset ExcelChart400K for training deep models on chart images. The code and the dataset are publicly available at https://github.com/soap117/DeepRule.
****

Shape From Semantic Segmentation via the Geometric Renyi Divergence
Tatsuro Koizumi, William A. P. Smith; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2312-2321
In this paper, we show how to estimate shape (restricted to a single object class via a 3D morphable model) using solely a semantic segmentation of a single 2D image. We propose a novel loss function based on a probabilistic, vertex-wise projection of the 3D model to the image plane. We represent both these projections and pixel labels as mixtures of Gaussians and compute the discrepancy between the two based on the geometric Renyi divergence. The resulting loss is differentiable and has a wide basin of convergence. We propose both classical, direct optimisation of this loss ("analysis-by-synthesis") and its use for training a parameter regression CNN. We show significant advantages over existing segmentation losses used in state-of-the-art differentiable renderers Soft Rasterizer and Neural Mesh Renderer.
****

Unsupervised Multi-Target Domain Adaptation Through Knowledge Distillation
Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1339-1347
Unsupervised domain adaptation (UDA) seeks to alleviate the problem of domain shift between the distribution of unlabeled data from the target domain w.r.t. labeled data from the source domain. While the single-target UDA scenario is well studied in the literature, Multi-Target Domain Adaptation (MTDA) remains largely unexplored despite its practical importance, e.g., in multi-camera video-surveillance applications. The MTDA problem can be addressed by adapting one specialized model per target domain, although this solution is too costly in many real-world applications. Blending multiple targets for MTDA has been proposed, yet this solution may lead to a reduction in model specificity and accuracy. In this paper, we propose a novel unsupervised MTDA approach to train a CNN that can general

ize well across multiple target domains. Our Multi-Teacher MTDA (MT-MTDA) method relies on multi-teacher knowledge distillation (KD) to iteratively distill target domain knowledge from multiple teachers to a common student. The KD process is performed in a progressive manner, where the student is trained by each teacher on how to perform UDA for a specific target, instead of directly learning domain adapted features. Finally, instead of combining the knowledge from each teacher, MT-MTDA alternates between teachers that distill knowledge, thereby preserving the specificity of each target (teacher) when learning to adapt to the student. MT-MTDA is compared against state-of-the-art methods on several challenging UDA benchmarks, and empirical results show that our proposed model can provide a considerably higher level of accuracy across multiple target domains. Our code is available at: https://github.com/LIVIAETS/MT-MTDA

****