Does Data Repair Lead to Fair Models? Curating Contextually Fair Data To Reduce Model Bias

Sharat Agarwal, Sumanyu Muku, Saket Anand, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3298-3307

Contextual information is a valuable cue for Deep Neural Networks (DNNs) to learn better representations and improve accuracy. However, co-occurrence bias in the training dataset may hamper a DNN model's generalizability to unseen scenarios in the real world. For example, in COCO [??], many object categories have a much higher co-occurrence with men compared to women, which can bias a DNN's prediction in favor of men. Recent works have focused on task-specific training strategies to handle bias in such scenarios, but fixing the available data is often ignored. In this paper, we propose a novel and more generic solution to address the contextual bias in the datasets by selecting a subset of the samples, which is fair in terms of the co-occurrence with various classes for a protected attribute. We introduce a data repair algorithm using the coefficient of variation(c_v), which can curate fair and contextually balanced data for a protected class(es). This helps in training a fair model irrespective of the task, architecture or training methodology. Our proposed solution is simple, effective and can even be used in an active learning setting where the data labels are not present or being generated incrementally. We demonstrate the effectiveness of our algorithm for the task of object detection and multi-label image classification across different datasets. Through a series of experiments, we validate that curating contextually fair data helps make model predictions fair by balancing the true positive rate for the protected class across groups without compromising on the model's overall performance.

****

UNETR: Transformers for 3D Medical Image Segmentation

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, Daguang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 574-584

Fully Convolutional Neural Networks (FCNNs) with contracting and expanding paths have shown prominence for the majority of medical image segmentation applications since the past decade. In FCNNs, the encoder plays an integral role by learning both global and local features and contextual representations which can be utilized for semantic output prediction by the decoder. Despite their success, the locality of convolutional layers in FCNNs, limits the capability of learning long-range spatial dependencies. Inspired by the recent success of transformers for Natural Language Processing (NLP) in long-range sequence learning, we reformulate the task of volumetric (3D) medical image segmentation as a sequence-to-sequence prediction problem. We introduce a novel architecture, dubbed as UNEt TRansformers (UNETR), that utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, while also following the successful "U-shaped" network design for the encoder and decoder. The transformer encoder is directly connected to a decoder via skip connections at different resolutions to compute the final semantic segmentation output. We have validated the performance of our method on the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) dataset for multi-organ segmentation and the Medical Segmentation Decathlon (MSD) dataset for brain tumor and spleen segmentation tasks. Our benchmarks demonstrate new state-of-the-art performance on the BTCV leaderboard.

****

SIDE: Center-Based Stereo 3D Detector With Structure-Aware Instance Depth Estimation

Xidong Peng, Xinge Zhu, Tai Wang, Yuexin Ma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 119-128

3D detection plays an indispensable role in environment perception. Due to the high cost of commonly used LiDAR sensor, stereo vision based 3D detection, as an economical yet effective setting, attracts more attention recently. For these approaches based on 2D images, accurate depth information is the key to achieve 3D detection, and most existing methods resort to a preliminary stage for depth estimation. They mainly focus on the global depth and neglect the property of depth information in this specific task, namely, sparsity and locality, where exactly accurate depth is only needed for these 3D bounding boxes. Motivated by this finding, we propose a stereo-image based anchor-free 3D detection method, called structure-aware stereo 3D detector (termed as SIDE), where we explore the instance-level depth information via constructing the cost volume from RoIs of each object. Due to the information sparsity of local cost volume, we further introduce match reweighting and structure-aware attention, to make the depth information more concentrated. Experiments conducted on the KITTI dataset show that our method achieves the state-of-the-art performance compared to existing methods without depth map supervision.

****

GraN-GAN: Piecewise Gradient Normalization for Generative Adversarial Networks

Vineeth S. Bhaskara, Tristan Aumentado-Armstrong, Allan D. Jepson, Alex Levinshtein; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3821-3830

Modern generative adversarial networks (GANs) predominantly use piecewise linear activation functions in discriminators (or critics), including ReLU and LeakyReLU. Such models learn piecewise linear mappings, where each piece handles a subset of the input space, and the gradients per subset are piecewise constant. Under such a class of discriminator (or critic) functions, we present Gradient Normalization (GraN), a novel input-dependent normalization method, which guarantees a piecewise K-Lipschitz constraint in the input space. In contrast to spectral normalization, GraN does not constrain processing at the individual network layers, and, unlike gradient penalties, strictly enforces a piecewise Lipschitz constraint almost everywhere. Empirically, we demonstrate improved image generation performance across multiple datasets (incl. CIFAR-10/100, STL-10, LSUN bedrooms, and CelebA), GAN loss functions, and metrics. Further, we analyze altering the often untuned Lipschitz constant K in several standard GANs, not only attaining significant performance gains, but also finding connections between K and training dynamics, particularly in low-gradient loss plateaus, with the common Adam optimizer.

****

Meta-UDA: Unsupervised Domain Adaptive Thermal Object Detection Using Meta-Learning

Vibashan VS, Domenick Poster, Suya You, Shuowen Hu, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1412-1423

Object detectors trained on large-scale RGB datasets are being extensively employed in real-world applications. However, these RGB-trained models suffer a performance drop under adverse illumination and lighting conditions. Infrared (IR) cameras are robust under such conditions and can be helpful in real-world applications. Though thermal cameras are widely used for military applications and increasingly for commercial applications, there is a lack of robust algorithms to robustly exploit the thermal imagery due to the limited availability of labeled thermal data. In this work, we aim to enhance the object detection performance in the thermal domain by leveraging the labeled visible domain data in an Unsupervised Domain Adaptation (UDA) setting. We propose an algorithm agnostic meta-learning framework to improve existing UDA methods instead of proposing a new UDA strategy. We achieve this by meta-learning the initial condition of the detector, which facilitates the adaptation process with fine updates without overfitting or getting stuck at local optima. However, meta-learning the initial condition for the detection scenario is computationally heavy due to long and intractable computation graphs. Therefore, we propose an online meta-learning paradigm which performs online updates resulting in a short and tractable computation graph. To this end, we demonstrate the superiority of our method over many baselines in the UDA setting, producing a state-of-the-art thermal detector for the KAIST and DSIAC datasets. Source code will be made publicly available after the review process.

****

Multi-Level Attentive Adversarial Learning With Temporal Dilation for Unsupervised Video Domain Adaptation

Peipeng Chen, Yuan Gao, Andy J. Ma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1259-1268

Most existing works on unsupervised video domain adaptation attempt to mitigate the distribution gap across domains in frame and video levels. Such two-level distribution alignment approach may suffer from the problems of insufficient alignment for complex video data and misalignment along the temporal dimension. To address these issues, we develop a novel framework of Multi-level Attentive Adversarial Learning with Temporal Dilation (MA2L-TD). Given frame-level features as input, multi-level temporal features are generated and multiple domain discriminators are individually trained by adversarial learning for them. For better distribution alignment, level-wise attention weights are calculated by the degree of domain confusion in each level. To mitigate the negative effect of misalignment, features are aggregated with the attention mechanism determined by individual domain discriminators. Moreover, temporal dilation is designed for sequential non-repeatability to balance the computational efficiency and the possible number of levels. Extensive experimental results show that our proposed method outperforms the state of the arts on four benchmark datasets.

****

Generative Adversarial Graph Convolutional Networks for Human Action Synthesis

Bruno Degardin, João Neves, Vasco Lopes, João Brito, Ehsan Yaghoubi, Hugo Proença; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1150-1159

Synthesising the spatial and temporal dynamics of the human body skeleton remains a challenging task, not only in terms of the quality of the generated shapes, but also of their diversity, particularly to synthesise realistic body movements of a specific action (action conditioning). In this paper, we propose Kinetic-GAN, a novel architecture that leverages the benefits of Generative Adversarial Networks and Graph Convolutional Networks to synthesise the kinetics of the human body. The proposed adversarial architecture can condition up to 120 different actions over local and global body movements while improving sample quality and diversity through latent space disentanglement and stochastic variations. Our experiments were carried out in three well-known datasets, where Kinetic-GAN notably surpasses the state-of-the-art methods in terms of distribution quality metrics while having the ability to synthesise more than one order of magnitude regarding the number of different actions. Our code and models are publicly available at https://github.com/DegardinBruno/Kinetic-GAN.

****

Non-Blind Deblurring for Fluorescence: A Deformable Latent Space Approach With Kernel Parameterization

Ziqiao Guan, Esther H. R. Tsai, Xiaojing Huang, Kevin G. Yager, Hong Qin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 711-719

Non-blind deblurring (NBD) is a modeling method of the image deblurring problem in computer vision, where the blurring kernel is known or can be externally estimated. In this paper, we attempt to solve a parametric NBD problem, inspired by the simultaneous acquisition of ptychography and fluorescent imaging (FI). Ptychography is an imaging method that favors larger probes, i.e. convolutional kernels, while FI relies on a small probe for high resolution. Also, the kernel can be solved during ptychographic reconstruction. With Ptycho-FI using the same larger kernel, we can perform NBD on the blurred fluorescent images to achieve high-resolution FI, and thus speed up the experiments. To this end, we design a deep latent space deformation network that is directly parameterized by the kernel. The network consists of three components: encoder, deformer, and decoder, where the deformer is specifically meant to rectify the latent space representations of blurred images to a standard latent space, regardless of the kernel. The deformation network is trained with a two-stage training scheme. We conduct extensive experiments to confirm that our parametric model can adapt to drastically differ ent blurring kernels and perform robust deblurring.

****

Few-Shot Open-Set Recognition of Hyperspectral Images With Outlier Calibration Network

Debabrata Pal, Valay Bundele, Renuka Sharma, Biplab Banerjee, Yogananda Jeppu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3801-3810

We tackle the few-shot open-set recognition (FSOSR) problem in the context of remote sensing hyperspectral image (HSI) classification. Prior research on OSR mainly considers an empirical threshold on the class prediction scores to reject the outlier samples. Further, recent endeavors in few-shot HSI classification fail to recognize outliers due to the `closed-set' nature of the problem and the fact that the entire class distributions are unknown during training. To this end, we propose to optimize a novel outlier calibration network (OCN) together with a feature extraction module during the meta-training phase. The feature extractor is equipped with a novel residual 3D convolutional block attention network (R3CBAM) for enhanced spectral-spatial feature learning from HSI. Our method rejects the outliers based on OCN prediction scores barring the need for manual thresholding. Finally, we propose to augment the query set with synthesized support set features during the similarity learning stage in order to combat the data scarcity issue of few-shot learning. The superiority of the proposed model is showcased on four benchmark HSI datasets.

****

SBEVNet: End-to-End Deep Stereo Layout Estimation

Divam Gupta, Wei Pu, Trenton Tabor, Jeff Schneider; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 523-532

Accurate layout estimation is crucial for planning and navigation in robotics applications, such as self-driving. In this paper, we introduce the Stereo Bird's Eye ViewNetwork (SBEVNet), a novel supervised end-to-end framework for estimation of bird's eye view layout from a pair of stereo images. Although our network reuses some of the building blocks from the state-of-the-art deep learning networks for disparity estimation, we show that explicit depth estimation is neither sufficient nor necessary. Instead, the learning of a good internal bird's eye view feature representation is effective for layout estimation. Specifically, we first generate a disparity feature volume using the features of the stereo images and then project it to the bird's eye view coordinates. This gives us coarse-grained information about the scene structure. We also apply inverse perspective mapping (IPM) to map the input images and their features to the bird's eye view. This gives us fine-grained texture information. Concatenating IPM features with the projected feature volume creates a rich bird's eye view representation which is useful for spatial reasoning. We use this representation to estimate the BEV semantic map. Additionally, we show that using the IPM features as a supervisory signal for stereo features can give an improvement in performance. We demonstrate our approach on two datasets:the KITTI dataset and a synthetically generated dataset from the CARLA simulator. For both of these datasets, we establish state-of-the-art performance compared to baseline techniques.

****

F-CAM: Full Resolution Class Activation Maps via Guided Parametric Upscaling

Soufiane Belharbi, Aydin Sarraf, Marco Pedersoli, Ismail Ben Ayed, Luke McCaffrey, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3490-3499

Class Activation Mapping (CAM) methods have recently gained much attention for weakly-supervised object localization (WSOL) tasks. They allow for CNN visualization and interpretation without training on fully annotated image datasets. CAM methods are typically integrated within off-the-shelf CNN backbones, such as ResNet50. Due to convolution and pooling operations, these backbones yield low resolution CAMs with a down-scaling factor of up to 32, contributing to inaccurate localizations. Interpolation is required to restore full size CAMs, yet it does not consider the statistical properties of objects, such as color and texture, leading to activations with inconsistent boundaries, and inaccurate localizations.

As an alternative, we introduce a generic method for parametric upscaling of CAMs that allows constructing accurate full resolution CAMs (F-CAMs). In particular, we propose a trainable decoding architecture that can be connected to any CNN classifier to produce highly accurate CAM localizations. Given an original low resolution CAM, foreground and background pixels are randomly sampled to fine-tune the decoder. Additional priors such as image statistics and size constraints are also considered to expand and refine object boundaries. Extensive experiments, over three CNN backbones and six WSOL baselines on the CUB-200-2011 and OpenImages datasets, indicate that our F-CAM method yields a significant improvement in CAM localization accuracy. F-CAM performance is competitive with state-of-art WSOL methods, yet it requires fewer computations during inference.

****

Auditing Saliency Cropping Algorithms

Abeba Birhane, Vinay Uday Prabhu, John Whaley; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4051-4059

In this paper, we audit saliency cropping algorithms used by Twitter, Google and Apple to investigate issues pertaining to the male-gaze cropping phenomenon as well as race-gender biases that emerge in post-cropping survival ratios of face-images constituting 3 x 1 grid images. In doing so, we present the first formal empirical study which suggests that the worry of a male-gaze-like image cropping phenomenon on Twitter is not at all far-fetched and it does occur with worryingly high prevalence rates in real-world full-body single-female-subject images shot with logo-littered backdrops. We uncover that while all three saliency cropping frameworks considered in this paper do exhibit acute racial and gender biases, Twitter's saliency cropping framework uniquely elicits high male-gaze cropping prevalence rates. In order to facilitate reproducing the results presented here, we are open-sourcing both the code and the datasets that we curated at shorturl.at/iuzK9. We hope the computer vision community and saliency cropping researchers will build on the results presented here and extend these investigations to similar frameworks deployed in the real world by other companies such as Microsoft and Facebook.

****

The Untapped Potential of Off-the-Shelf Convolutional Neural Networks

Matthew Inkawhich, Nathan Inkawhich, Eric Davis, Hai Li, Yiran Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2818-2827

Over recent years, a myriad of novel convolutional network architectures have been developed to advance state-of-the-art performance on challenging recognition tasks. As computational resources improve, a great deal of effort has been placed on efficiently scaling up existing designs and generating new architectures with Neural Architecture Search (NAS) algorithms. While network topology has proven to be a critical factor for model performance, we show that significant gains are being left on the table by keeping topology static at inference-time. Due to challenges such as scale variation, we should not expect static models configured to perform well across a training dataset to be optimally configured to handle all test data. In this work, we expose the exciting potential of inference-time-dynamic models. We show that by allowing just four layers to dynamically change configuration at inference-time, off-the-shelf models like ResNet-50 have an upper bound accuracy of over 95% on ImageNet. This level of performance currently exceeds that of models with over 20x more parameters and significantly more complex training procedures. While this upper bound of performance may be practically difficult to achieve for a real dynamic model, it indicates a significant source of untapped potential for current models.

****

X-MIR: EXplainable Medical Image Retrieval

Brian Hu, Bhavan Vasu, Anthony Hoogs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 440-450

Despite significant progress in the past few years, machine learning systems are still often viewed as "black boxes", which lack the ability to explain their output decisions. In high-stakes situations such as healthcare, there is a need fo r explainable AI (XAI) tools that can help open up this black box. In contrast to approaches which largely tackle classification problems in the medical imaging domain, we address the less-studied problem of explainable image retrieval. We test our approach on a COVID-19 chest X-ray dataset and the ISIC 2017 skin lesion dataset, showing that saliency maps help reveal the image features used by models to determine image similarity. We evaluated three different saliency algorithms, which were either occlusion-based, attention-based, or relied on a form of activation mapping. We also develop quantitative evaluation metrics that allow us to go beyond simple qualitative comparisons of the different saliency algorithms. Our results have the potential to aid clinicians when viewing medical images and addresses an urgent need for interventional tools in response to COVID-19. The source code is publicly available at: https://gitlab.kitware.com/brianhhu/x-mir.

****

On the Effectiveness of Small Input Noise for Defending Against Query-Based Black-Box Attacks

Junyoung Byun, Hyojun Go, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3051-3060

While deep neural networks show unprecedented performance in various tasks, the vulnerability to adversarial examples hinders their deployment in safety-critical systems. Many studies have shown that attacks are also possible even in a black-box setting where an adversary cannot access the target model's internal information. Most black-box attacks are based on queries, each of which obtains the target model's output for an input, and many recent studies focus on reducing the number of required queries. In this paper, we pay attention to an implicit assumption of query-based black-box adversarial attacks that the target model's output exactly corresponds to the query input. If some randomness is introduced into the model, it can break the assumption, and thus, query-based attacks may have tremendous difficulty in both gradient estimation and local search, which are the core of their attack process. From this motivation, we observe even a small additive input noise can neutralize most query-based attacks and name this simple yet effective approach Small Noise Defense (SND). We analyze how SND can defend against query-based black-box attacks and demonstrate its effectiveness against eight state-of-the-art attacks with CIFAR-10 and ImageNet datasets. Even with strong defense ability, SND almost maintains the original classification accuracy and computational speed. SND is readily applicable to pre-trained models by adding only one line of code at the inference.

****

A Fast Partial Video Copy Detection Using KNN and Global Feature Database

Weijun Tan, Hongwei Guo, Rushuai Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2191-2199

Unlike in most previous partial video copy detection (PVCD) algorithms, where reference videos are scanned one by one, we treat the PVCD as a video search/retrieval problem. We propose a fast partial video copy detection framework in this paper. In this framework, all frame CNN features of the reference videos are organized in a KNN searchable database. Instead of scanning all reference videos, the query video segment does a fast KNN search in the global feature database. The returned results are used to generate a shortlist of candidate videos. A modified temporal network is then used to localize the copy segment in the candidate videos. Furthermore, We propose to use a transformer encoder to improve the CNN feature. We evaluate our algorithm on the VCDB dataset. Our benchmark F1 scores exceed state-of-the-art by a big margin. The speed of our algorithm is also improved significantly.

****

Information Bottlenecked Variational Autoencoder for Disentangled 3D Facial Expression Modelling

Hao Sun, Nick Pears, Yajie Gu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 157-166

Learning a disentangled representation is essential to build 3D face models that accurately capture identity and expression. We propose a novel variational auto

encoder (VAE) framework to disentangle identity and expression from 3D input faces that have a wide variety of expressions. Specifically, we design a system that has two decoders: one for neutral-expression faces (i.e. identity-only faces) and one for the original (expressive) input faces respectively. Crucially, we have an additional mutual-information regulariser applied on the identity part to solve the issue of imbalanced information over the expressive input faces and the reconstructed neutral faces. Our evaluations on two public datasets (CoMA and BU-3DFE) show that this model achieves competitive results on the 3D face reconstruction task and state-of-the-art results on identity-expression disentanglement. We also show that by updating to a conditional VAE, we have a system that generates different levels of expressions from semantically meaningful variables.
****

Discrete Neural Representations for Explainable Anomaly Detection
Stanislaw Szymanowicz, James Charles, Roberto Cipolla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 148-156
The aim of this work is to detect and automatically generate high-level explanations of anomalous events in video. Understanding the cause of an anomalous event is crucial as the required response is dependant on its nature and severity. Recent works typically use object or action classifier to detect and provide labels for anomalous events. However, this constrains detection systems to a finite set of known classes and prevents generalisation to unknown objects or behaviours. Here we show how to robustly detect anomalies without the use of object or action classifiers yet still recover the high level reason behind the event. We make the following contributions: (1) a method using saliency maps to decouple the explanation of anomalous events from object and action classifiers, (2) show how to improve the quality of saliency maps using a novel neural architecture for learning discrete representations of video by predicting future frames and (3) beat the state-of-the-art anomaly explanation methods by 60% on a subset of the public benchmark X-MAN dataset.
****

Attack Agnostic Detection of Adversarial Examples via Random Subspace Analysis
Nathan Drenkow, Neil Fendley, Philippe Burlina; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 472-482
Whilst adversarial attack detection has received considerable attention, it remains a fundamentally challenging problem from two perspectives. First, while threat models can be well-defined, attacker strategies may still vary widely within those constraints. Therefore, detection should be considered as an open-set problem, standing in contrast to most current detection approaches. These methods take a closed-set view and train binary detectors, thus biasing detection toward attacks seen during detector training. Second, limited information is available at test time and typically confounded by nuisance factors including the label and underlying content of the image. We address these challenges via a novel strategy based on random subspace analysis. We present a technique that utilizes properties of random projections to characterize the behavior of clean and adversarial examples across a diverse set of subspaces. The self-consistency (or inconsistency) of model activations is leveraged to discern clean from adversarial examples. Performance evaluations demonstrate that our technique (AUC [0.92, 0.98]) outperforms competing detection strategies (AUC [0.30,0.79]), while remaining truly agnostic to the attack strategy (for both targeted/untargeted attacks). It also requires significantly less calibration data (composed only of clean examples) than competing approaches to achieve this performance.
****

Distance-Based Hyperspherical Classification for Multi-Source Open-Set Domain Adaptation
Silvia Bucci, Francesco Cappio Borlino, Barbara Caputo, Tatiana Tommasi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1119-1128
Vision systems trained in closed-world scenarios fail when presented with new environmental conditions, new data distributions, and novel classes at deployment

time. How to move towards open-world learning is a long-standing research question. The existing solutions mainly focus on specific aspects of the problem (single domain Open-Set, multi-domain Closed-Set), or propose complex strategies which combine several losses and manually tuned hyperparameters. In this work, we tackle multi-source Open-Set domain adaptation by introducing HyMOS: a straightforward model that exploits the power of contrastive learning and the properties of its hyperspherical feature space to correctly predict known labels on the target, while rejecting samples belonging to any unknown class. HyMOS includes style transfer among the instance transformations of contrastive learning to get domain invariance while avoiding the risk of negative-transfer. A self-paced threshold is defined on the basis of the observed data distribution and updates online during training, allowing to handle the known-unknown separation. We validate our method over three challenging datasets. The obtained results show that HyMOS outperforms several competitors, defining the new state-of-the-art. Our code is available at https://github.com/silvia1993/HyMOS.

****

D2Conv3D: Dynamic Dilated Convolutions for Object Segmentation in Videos
Christian Schmidt, Ali Athar, Sabarinath Mahadevan, Bastian Leibe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1200-1209

Despite receiving significant attention from the research community, the task of segmenting and tracking objects in monocular videos still has much room for improvement. Existing works have simultaneously justified the efficacy of dilated and deformable convolutions for various image-level segmentation tasks. This gives reason to believe that 3D extensions of such convolutions should also yield performance improvements for video-level segmentation tasks. However, this aspect has not yet been explored thoroughly in existing literature. In this paper, we propose Dynamic Dilated Convolutions (D2Conv3D): a novel type of convolution which draws inspiration from dilated and deformable convolutions and extends them to the 3D (spatio-temporal) domain. We experimentally show that D2Conv3D can be used to improve the performance of multiple 3D CNN architectures across multiple video segmentation related benchmarks by simply employing D2Conv3D as a drop-in replacement for standard convolutions. We further show that D2Conv3D out-performs trivial extensions of existing dilated and deformable convolutions to 3D. Lastly, we set a new state-of-the-art on the DAVIS 2016 Unsupervised Video Object Segmentation benchmark. Code is made publicly available at https://github.com/Schmiddo/d2conv3d.

****

Multi-Motion and Appearance Self-Supervised Moving Object Detection
Fan Yang, Srikrishna Karanam, Meng Zheng, Terrence Chen, Haibin Ling, Ziyan Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2605-2614

In this work, we consider the problem of self-supervised Moving Object Detection (MOD) in video, where no ground truth is involved in both training and inference phases. Recently, an adversarial learning framework is proposed to leverage inherent temporal information for MOD. While showing great promising results, it uses single scale temporal information and may meet problems when dealing with a deformable object under multi-scale motion in different parts. Additional challenges can arise from the moving camera, which results in the failure of the motion independence hypothesis and locally independent background motion. To deal with these problems, we propose a Multi-motion and Appearance Self-supervised Network (MASNet) to introduce multi-scale motion information and appearance information of scene for MOD. In particular, a moving object, especially the deformable, usually consists of moving regions at various temporal scales. Introducing multi-scale motion can aggregate these regions to form a more complete detection. Appearance information can serve as another cue for MOD when the motion independence is not reliable and for removing false detection in background caused by locally independent background motion. To encode multi-scale motion and appearance, in MASNet we respectively design a multi-branch flow encoding module and an image inpainter module. The proposed modules and MASNet are extensively evaluated on the DAVIS dataset to demonstrate the effectiveness and superiority to state-of-the-art self-supervised methods.

****

CeyMo: See More on Roads - A Novel Benchmark Dataset for Road Marking Detection

Oshada Jayasinghe, Sahan Hemachandra, Damith Anhettigama, Shenali Kariyawasam, Ranga Rodrigo, Peshala Jayasekara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3104-3113

In this paper, we introduce a novel road marking benchmark dataset for road marking detection, addressing the limitations in the existing publicly available datasets such as lack of challenging scenarios, prominence given to lane markings, unavailability of an evaluation script, lack of annotation formats and lower resolutions. Our dataset consists of 2887 total images with 4706 road marking instances belonging to 11 classes. The images have a high resolution of 1920 x 1080 and capture a wide range of traffic, lighting and weather conditions. We provide road marking annotations in polygons, bounding boxes and pixel-level segmentation masks to facilitate a diverse range of road marking detection algorithms. The evaluation metrics and the evaluation script we provide, will further promote direct comparison of novel approaches for road marking detection with existing methods. Furthermore, we evaluate the effectiveness of using both instance segmentation and object detection based approaches for the road marking detection task. Speed and accuracy scores for two instance segmentation models and two object detector models are provided as a performance baseline for our benchmark dataset. The dataset and the evaluation script is publicly available.

****

Pixel-Level Bijective Matching for Video Object Segmentation

Suhwan Cho, Heansung Lee, Minjung Kim, Sungjun Jang, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 129-138

Semi-supervised video object segmentation (VOS) aims to track a designated object present in the initial frame of a video at the pixel level. To fully exploit the appearance information of an object, pixel-level feature matching is widely used in VOS. Conventional feature matching runs in a surjective manner, i.e., only the best matches from the query frame to the reference frame are considered. Each location in the query frame refers to the optimal location in the reference frame regardless of how often each reference frame location is referenced. This works well in most cases and is robust against rapid appearance variations, but may cause critical errors when the query frame contains background distractors that look similar to the target object. To mitigate this concern, we introduce a bijective matching mechanism to find the best matches from the query frame to the reference frame and vice versa. Before finding the best matches for the query frame pixels, the optimal matches for the reference frame pixels are first considered to prevent each reference frame pixel from being overly referenced. As this mechanism operates in a strict manner, i.e., pixels are connected if and only if they are the sure matches for each other, it can effectively eliminate background distractors. In addition, we propose a mask embedding module to improve the existing mask propagation method. By utilizing multiple historic masks and their variations, it can effectively capture the position information of a target object.

****

Unveiling Real-Life Effects of Online Photo Sharing

Van-Khoa Nguyen, Adrian Popescu, Jérôme Deshayes-Chossart; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2898-2908

Social networks give free access to their services in exchange for the right to exploit their users' data. Data sharing is done in an initial context which is chosen by the users. However, data are used by social networks and third parties in different contexts which are often not transparent. In order to unveil such usages, we propose an approach which focuses on the effects of data sharing in impactful real-life situations. Focus is put on visual content because of its strong influence in shaping online user profiles. The approach relies on three compo nents: (1) a set of concepts with associated situation impact ratings obtained by crowdsourcing, (2) a corresponding set of object detectors for mining users' photos and (3) a ground truth dataset made of 500 visual user profiles which are manually rated per situation. These components are combined in LERVUP, a method which learns to rate visual user profiles in each situation. LERVUP exploits a new image descriptor which aggregates concept ratings and object detections at user level and an attention mechanism which boosts highly-rated concepts to prevent them from being overwhelmed by low-rated ones. Performance is evaluated per situation by measuring the correlation between the automatic ranking of profile ratings and a manual ground truth. Results indicate that LERVUP is effective since a strong correlation of the two rankings is obtained. A practical implementation of the approach in a mobile app which raises user awareness about shared data usage is also discussed.

****

HierMatch: Leveraging Label Hierarchies for Improving Semi-Supervised Learning

Ashima Garg, Shaurya Bagga, Yashvardhan Singh, Saket Anand; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1015-1024

Semi-supervised learning approaches have emerged as an active area of research to combat the challenge of obtaining large amounts of annotated data. Towards the goal of improving the performance of semi-supervised learning methods, we propose a novel framework, HIERMATCH, a semi-supervised approach that leverages hierarchical information to reduce labeling costs and performs as well as a vanilla semi-supervised learning method. Hierarchical information is often available as prior knowledge in the form of coarse labels (e.g., woodpeckers) for images with fine-grained labels (e.g., downy woodpeckers or golden-fronted woodpeckers). However, the use of supervision using coarse-category labels to improve semi-supervised techniques has not been explored. In the absence of fine-grained labels, HIERMATCH exploits the label hierarchy and uses coarse class labels as a weak supervisory signal. Additionally, HIERMATCH is a generic-approach to improve any semi-supervised learning framework, we demonstrate this using our results on recent state-of-the-art techniques MixMatch and FixMatch. We evaluate the efficacy of HIERMATCH on two benchmark datasets, namely CIFAR-100 and NABirds. HIERMATCH can reduce the usage of fine-grained labels by 50% on CIFAR-100 with only a marginal drop of 0.59% in top-1 accuracy as compared to MixMatch.

****

Mobile Based Human Identification Using Forehead Creases: Application and Assessment Under COVID-19 Masked Face Scenarios

Rohit Bharadwaj, Gaurav Jaswal, Aditya Nigam, Kamlesh Tiwari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3693-3701

In the COVID-19 situation, face masks have become an essential part of our daily life. As mask occludes most prominent facial characteristics, it brings new challenges to the existing facial recognition systems. This paper presents an idea to consider forehead creases (under surprise facial expression) as a new biometric modality to authenticate mask-wearing faces. The forehead biometrics utilizes the creases and textural skin patterns appearing due to voluntary contraction of the forehead region as features. The proposed framework is an efficient and generalizable deep learning framework for forehead recognition. Face-selfie images are collected using smartphone's frontal camera in an unconstrained environment with various indoor/outdoor realistic environments. Acquired forehead images are first subjected to a segmentation model that results in rectangular Region Of Interest (ROI's). A set of convolutional feature maps are subsequently obtained using a backbone network. The primary embeddings are enriched using a dual attention network (DANet) to induce discriminative feature learning. The attention-empowered embeddings are then optimized using Large Margin Cosine Loss (LMCL) followed by Focal Loss to update weights for inducting robust training and better feature discriminating capabilities. Our system is end-to-end and few-shot; thus, it is very efficient in memory requirements and recognition rate. Besides, we present a forehead image dataset that has been recorded in two sessions from 247 s ubjects containing a total of 4,964 selfie-face mask images. To the best of our knowledge, this is the first to date mobile-based forehead dataset and is being made available along with the mobile application in the public domain. The proposed system has achieved high performance results in both closed-set, i.e., CRR of 99.08% and EER of 0.44% and open-set matching, i.e., CRR: 97.84%, EER: 12.40% which justifies the significance of using forehead as a biometric modality.

****

Beyond Mono to Binaural: Generating Binaural Audio From Mono Audio With Depth and Cross Modal Attention

Kranti Kumar Parida, Siddharth Srivastava, Gaurav Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3347-3356

Binaural audio gives the listener an immersive experience and can enhance augmented and virtual reality. However, recording binaural audio requires specialized setup with a dummy human head having microphones in left and right ears. Such a recording setup is difficult to build and setup, therefore mono audio has become the preferred choice in common devices. To obtain the same impact as binaural audio, recent efforts have been directed towards lifting mono audio to binaural audio conditioned on the visual input from the scene. Such approaches have not used an important cue for the task: the distance of different sound producing objects from the microphones. In this work, we argue that depth map of the scene can act as a proxy for inducing distance information of different objects in the scene, for the task of audio binauralization. We propose a novel encoder-decoder architecture with a hierarchical attention mechanism to encode image, depth and audio feature jointly. We design the network on top of state-of-the-art transformer networks for image and depth representation. We show empirically that the proposed method outperforms state-of-the-art methods comfortably for two challenging public datasets FAIR-Play and MUSIC- Stereo. We also demonstrate with qualitative results that the method is able to focus on the right information required for the task. The qualitative results are available at our project page https://krantiparida.github.io/projects/bmonobinaural.html

****

Matching and Recovering 3D People From Multiple Views

Alejandro Perez-Yus, Antonio Agudo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3622-3631

This paper introduces an approach to simultaneously match and recover 3D people from multiple calibrated cameras. To this end, we present an affinity measure between 2D detections across different views that enforces an uncertainty geometric consistency. This similarity is then exploited by a novel multi-view matching algorithm to cluster the detections, being robust against partial observations as well as bad detections and without assuming any prior about the number of people in the scene. After that, the multi-view correspondences are used in order to efficiently infer the 3D pose of each body by means of a 3D pictorial structure model in combination with physico-geometric constraints. Our algorithm is thoroughly evaluated on challenging scenarios where several human bodies are performing different activities which involve complex motions, producing large occlusions in some views and noisy observations. We outperform state-of-the-art results in terms of matching and 3D reconstruction.

****

Rethinking Video Anomaly Detection - A Continual Learning Approach

Keval Doshi, Yasin Yilmaz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3961-3970

While video anomaly detection has been an active area of research for several years, recent progress is limited to improving the state-of-the-art results on small datasets using an inadequate evaluation criterion. In this work, we take a new comprehensive look at the video anomaly detection problem from a more realistic perspective. Specifically, we consider practical challenges such as continual learning and few-shot learning, which humans can easily do but remains to be a significant challenge for machines. A novel algorithm designed for such practical challenges is also proposed. For performance evaluation in this new framework,

we introduce a new dataset which is significantly more comprehensive than the existing benchmark datasets, and a new performance metric which takes into account the fundamental temporal aspect of video anomaly detection. The experimental results show that the existing state-of-the-art methods are not suitable for the considered practical challenges, and the proposed algorithm outperforms them with a large margin in continual learning and few-shot learning tasks.
****

A Pixel-Level Meta-Learner for Weakly Supervised Few-Shot Semantic Segmentation
Yuan-Hao Lee, Fu-En Yang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2170-2180
Few-shot semantic segmentation addresses the learning task in which only few images with ground truth pixel-level labels are available for the novel classes of interest. One is typically required to collect a large mount of data (i.e., base classes) with such ground truth information, followed by meta-learning strategies to address the above learning task. When only image-level semantic labels can be observed during both training and testing, it is considered as an even more challenging task of weakly supervised few-shot semantic segmentation. To address this problem, we propose a novel meta-learning framework, which predicts pseudo pixel-level segmentation masks from a limited amount of data and their semantic labels. More importantly, our learning scheme further exploits the produced pixel-level information for query image inputs with segmentation guarantees. Thus, our proposed learning model can be viewed as a pixel-level meta-learner. Through extensive experiments on benchmark datasets, we show that our model achieves satisfactory performances under fully supervised settings, yet performs favorably against state-of-the-art methods under weakly supervised settings.
****

Plugging Self-Supervised Monocular Depth Into Unsupervised Domain Adaptation for Semantic Segmentation
Adriano Cardace, Luca De Luigi, Pierluigi Zama Ramirez, Samuele Salti, Luigi Di Stefano; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1129-1139
Although recent semantic segmentation methods have made remarkable progress, they still rely on large amounts of annotated training data, which are often infeasible to collect in the autonomous driving scenario. Previous works usually tackle this issue with Unsupervised Domain Adaptation (UDA), which entails training a network on synthetic images and applying the model to real ones while minimizing the discrepancy between the two domains. Yet, these techniques do not consider additional information that may be obtained from other tasks. Differently, we propose to exploit self-supervised monocular depth estimation to improve UDA for semantic segmentation. On one hand, we deploy depth to realize a plug-in component which can inject complementary geometric cues into any existing UDA method. We further rely on depth to generate a large and varied set of samples to Self-Train the final model. Our whole proposal allows for achieving state-of-the-art performance (58.8 mIoU) in the GTA5->CS benchmark benchmark. Code is available at https://github.com/CVLAB-Unibo/d4-dbst.
****

Training a Task-Specific Image Reconstruction Loss
Aamir Mustafa, Aliaksei Mikhailiuk, Dan Andrei Iliescu, Varun Babbar, Rafa■ K. Mantiuk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2319-2328
The choice of a loss function is an important factor when training neural networks for image restoration problems, such as single image super resolution. The loss function should encourage natural and perceptually pleasing results. A popular choice for a loss is a pre-trained network, such as VGG, which is used as a feature extractor for computing the difference between restored and reference images. However, such an approach has multiple drawbacks: it is computationally expensive, requires regularization and hyper-parameter tuning, and involves a large network trained on an unrelated task. Furthermore, it has been observed that there is no single loss function that works best across all applications and across different datasets. In this work, we instead propose to train a set of loss fun

ctions that are application specific in nature. Our loss function comprises a series of discriminators that are trained to detect and penalize the presence of application-specific artefacts. We show that a single natural image and corresponding distortions are sufficient to train our feature extractor that outperforms state-of-the-art loss functions in applications like single image super resolution, denoising, and JPEG artefact removal. Finally, we conclude that an effective loss function does not have to be a good predictor of perceived image quality, but instead needs to be specialized in identifying the distortions for a given restoration method.

****

Learning to Weight Filter Groups for Robust Classification

Siyang Yuan, Yitong Li, Dong Wang, Ke Bai, Lawrence Carin, David Carlson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3041-3050

In many real-world tasks, a canonical "big data" problem is created by combining data from several individual groups or domains. Because test data will likely come from a new group of data, we want to utilize the grouped structure of our training data to enforce generalization between groups of data, not just individual samples. This can be viewed as a multiple-domain generalization problem. Specifically, the goal is to encourage generalization between previously seen labeled source data from multiple domains and unlabeled target domain data. To address this challenge, we introduce Domain-Specific Filter Group (DSFG), where each training domain has a unique filter group and each test data point is predicted by a weighted sum over the outputs of different domain filters. A separate neural network learns to estimate the appropriate filter group weights through a meta-learning strategy. Empirically, experiments on three benchmark datasets demonstrate improved performance compared to current state-of-the-art approaches.

****

Adversarial Open Domain Adaptation for Sketch-to-Photo Synthesis

Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, Jan P. Allebach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1434-1444

In this paper, we explore open-domain sketch-to-photo translation, which aims to synthesize a realistic photo from a freehand sketch with its class label, even if the sketches of that class are missing in the training data. It is challenging due to the lack of training supervision and the large geometric distortion between the freehand sketch and photo domains. To synthesize the absent freehand sketches from photos, we propose a framework that jointly learns sketch-to-photo and photo-to-sketch generation. However, the generator trained from fake sketches might lead to unsatisfying results when dealing with sketches of missing classes, due to the domain gap between synthesized sketches and real ones. To alleviate this issue, we further propose a simple yet effective open-domain sampling and optimization strategy to "fool" the generator into treating fake sketches as real ones. Our method takes advantage of the learned sketch-to-photo and photo-to-sketch mapping of in-domain data and generalizes it to the open-domain classes. We validate our method on the Scribble and SketchyCOCO datasets. Compared with the recent competing methods, our approach shows impressive results in synthesizing realistic color, texture, and maintaining the geometric composition for various categories of open-domain sketches.

****

Shadow Art Revisited: A Differentiable Rendering Based Approach

Kaustubh Sadekar, Ashish Tiwari, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 29-37

While recent learning-based methods have been observed to be superior for several vision-related applications, their potential in generating artistic effects has not been explored much. One such exciting application is Shadow Art - a unique form of sculptural art that produces artistic effects through 2D shadows cast by a 3D sculpture. In this work, we revisit shadow art using differentiable rendering-based optimization frameworks to obtain the 3D sculpture from a set of shadow (binary) images and their corresponding projection information. Specifically, we discuss shape optimization through voxel as well as mesh-based differentiable renderers. Our choice of using differentiable rendering for generating shadow art sculptures can be attributed to its ability to learn the underlying 3D geometry solely from image data, thus reducing the dependence on 3D ground truth. The qualitative and quantitative results demonstrate the potential of the proposed framework in generating complex 3D sculptures that transcend the ones seen in contemporary art pieces using just a set of shadow images as input. Further, we demonstrate the generation of 3D sculptures to cast shadows of faces, animated movie characters, and the applicability of the proposed framework to sketch-based 3D reconstruction of the underlying shapes.

****

Occlusion Resistant Network for 3D Face Reconstruction

Hitika Tiwari, Vinod K. Kurmi, K.S. Venkatesh, Yong-Sheng Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 813-822

3D face reconstruction from a monocular face image is a mathematically ill-posed problem. Recently, we observed a surge of interest in deep learning-based approaches to address the issue. These methods possess extreme sensitivity towards occlusions. Thus, in this paper, we present a novel context-learning-based distillation approach to tackle the occlusions in the face images. Our training pipeline focuses on distilling the knowledge from a pre-trained occlusion-sensitive deep network. The proposed model learns the context of the target occluded face image. Hence our approach uses a weak model (unsuitable for occluded face images) to train a highly robust network towards partially and fully-occluded face images. We obtain a landmark accuracy of 0.77 against 5.84 of recent state-of-the-art-method for real-life challenging facial occlusions. Also, we propose a novel end-to-end training pipeline to reconstruct 3D faces from multiple variations of the target image per identity to emphasize the significance of visible facial features during learning. For this purpose, we leverage a novel composite multi-occlusion loss function. Our multi-occlusion per identity model shows a dip in the landmark error by a large margin of 6.67 in comparison to a recent state-of-the-art method. We deploy the occluded variations of the CelebA validation dataset and AFLW2000-3D face dataset: naturally-occluded and artificially occluded, for the comparisons. We comprehensively compare our results with the other approaches concerning the accuracy of the reconstructed 3D face mesh for occluded face images.

****

Image Restoration by Deep Projected GSURE

Shady Abu-Hussein, Tom Tirer, Se Young Chun, Yonina C. Eldar, Raja Giryes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3602-3611

Ill-posed inverse problems appear in many image processing applications, such as deblurring and super-resolution. In recent years, solutions that are based on deep Convolutional Neural Networks (CNNs) have shown great promise. Yet, most of these techniques, which train CNNs using external data, are restricted to the observation models that have been used in the training phase. A recent alternative that does not have this drawback relies on learning the target image using internal learning. One such prominent example is the Deep Image Prior (DIP) technique that trains a network directly on the input image with the least-squares loss. In this paper, we propose a new image restoration framework that is based on minimizing a loss function that includes a "projected-version" of the Generalized Stein Unbiased Risk Estimator (GSURE) and parameterization of the latent image by a CNN. We demonstrate two ways to use our framework. In the first one, where no explicit prior is used, we show that the proposed approach outperforms other internal learning methods, such as DIP. In the second one, we show that our GSURE-based loss leads to improved performance when used within a plug-and-play priors scheme.

****

Multimodal Learning Using Optimal Transport for Sarcasm and Humor Detection

Shraman Pramanick, Aniket Roy, Vishal M. Patel; Proceedings of the IEEE/CVF Wint er Conference on Applications of Computer Vision (WACV), 2022, pp. 3930-3940

Multimodal learning is an emerging yet challenging research area. In this paper, we deal with multimodal sarcasm and humor detection from conversational videos and image-text pairs. Being a fleeting action, which is dependent across the modalities, sarcasm detection is challenging since large datasets are not available for this task in the literature. Therefore, we primarily focus on resource-constrained training, where the number of training samples is limited. To this end, we propose a novel multimodal learning system, MuLOT (Multimodal Learning using Optimal Transport), which utilizes self-attention to exploit intra-modal correspondence and optimal transport for cross-modal correspondence. Finally, the modalities are combined with multimodal attention fusion to capture the inter-dependencies across modalities. We test our proposed approach for multimodal sarcasm and humor detection on three benchmark datasets - MUStARD (video, audio, text), UR-FUNNY (video, audio, text), MST (image, text) and obtain 2.1%, 1.54%, and 2.34% accuracy improvements over the state-of-the-art.

****

Normalizing Flow as a Flexible Fidelity Objective for Photo-Realistic Super-Resolution

Andreas Lugmayr, Martin Danelljan, Fisher Yu, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1756-1765

Super-resolution is an ill-posed problem, where a ground-truth high-resolution image represents only one possibility in the space of plausible solutions. Yet, the dominant paradigm is to employ pixel-wise losses, such as L_1, which drive the prediction towards a blurry average. This leads to fundamentally conflicting objectives when combined with adversarial losses, which degrades the final quality. We address this issue by revisiting the L_1 loss and show that it corresponds to a one-layer conditional flow. Inspired by this relation, we explore general flows as a fidelity-based alternative to the L_1 objective. We demonstrate that the flexibility of deeper flows leads to better visual quality and consistency when combined with adversarial losses. We conduct extensive user studies for three datasets and scale factors, where our approach is shown to outperform state-of-the-art methods for photo-realistic super-resolution.

****

Towards Class-Oriented Poisoning Attacks Against Neural Networks

Bingyin Zhao, Yingjie Lao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3741-3750

Poisoning attacks on machine learning systems compromise the model performance by deliberately injecting malicious samples in the training dataset to influence the training process. Prior works focus on either availability attacks (i.e., lowering the overall model accuracy) or integrity attacks (i.e., enabling specific instance-based backdoor). In this paper, we advance the adversarial objectives of the availability attacks to a per-class basis, which we refer to as class-oriented poisoning attacks. We demonstrate that the proposed attack is capable of forcing the corrupted model to predict in two specific ways: (i) classify unseen new images to a targeted "supplanter" class, and (ii) misclassify images from a "victim" class while maintaining the classification accuracy on other non-victim classes. To maximize the adversarial effect as well as reduce the computational complexity of poisoned data generation, we propose a gradient-based framework that crafts poisoning images with carefully manipulated feature information for each scenario. Using newly defined metrics at the class level, we demonstrate the effectiveness of the proposed class-oriented poisoning attacks on various models (e.g., LeNet-5, Vgg-9, and ResNet-50) over a wide range of datasets (e.g., MNIST, CIFAR-10, and ImageNet-ILSVRC2012) in an end-to-end training setting.

****

Transfer Learning for Pose Estimation of Illustrated Characters

Shuhong Chen, Matthias Zwicker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 793-802

Human pose information is a critical component in many downstream image processing tasks, such as activity recognition and motion tracking. Likewise, a pose est

imator for the illustrated character domain would provide a valuable prior for assistive content creation tasks, such as reference pose retrieval and automatic character animation. But while modern data-driven techniques have substantially improved pose estimation performance on natural images, little work has been done for illustrations. In our work, we bridge this domain gap by efficiently transfer-learning from both domain-specific and task-specific source models. Additionally, we upgrade and expand an existing illustrated pose estimation dataset, and introduce two new datasets for classification and segmentation subtasks. We then apply the resultant state-of-the-art character pose estimator to solve the novel task of pose-guided illustration retrieval. All data, models, and code will be made publicly available.
****
From Node To Graph: Joint Reasoning on Visual-Semantic Relational Graph for Zero-Shot Detection
Hui Nie, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1109-1118
Zero-Shot Detection (ZSD), which aims at localizing and recognizing unseen objects in a complicated scene, usually leverages the visual and semantic information of individual objects alone. However, scene understanding of human exceeds recognizing individual objects separately: the contextual information among multiple objects such as visual relational information (e.g. visually similar objects) and semantic relational information (e.g. co-occurrences) is helpful for understanding of visual scene. In this paper, we verify that contextual information plays a more important role in ZSD than in traditional object detection. To make full use of such information, we propose a new end-to-end ZSD method GRaph Aligning Network (GRAN) based on graph modeling and reasoning which simultaneously considers visual and semantic information of multiple objects instead of individual objects. Specifically, we formulate a Visual Relational Graph (VRG) and a Semantic Relational Graph (SRG), where the nodes are the objects in the image and the semantic representations of classes respectively and the edges are the relevance between nodes in each graph. To characterize mutual effect between two modalities, the two graphs are further merged into a heterogeneous Visual-Semantic Relational Graph (VSRG), where modal translators are designed for the two subgraphs to enable modal information to transform into a common space for communication, and message passing among nodes is enforced to refine their representations. Comprehensive experiments on MSCOCO dataset demonstrate the advantage of our method over state-of-the-arts, and qualitative analysis suggests the validity of using contextual information.
****
Unsupervised Sounding Object Localization With Bottom-Up and Top-Down Attention
Jiayin Shi, Chao Ma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1737-1746
Learning to localize sounding objects in visual scenes without manual annotations has drawn increasing attention recently. In this paper, we propose an unsupervised sounding object localization algorithm by using bottom-up and top-down attention in visual scenes. The bottom-up attention module generates an objectness confidence map, while the top-down attention draws the similarity between sound and visual regions. Moreover, we propose a bottom-up attention loss function, which models the correlation relationship between bottom-up and top-down attention. Extensive experimental results demonstrate that our proposed unsupervised method significantly advances the state-of-the-art unsupervised methods. The source code is available at https://github.com/VISION-SJTU/usol/.
****
Multi-Scale Patch-Based Representation Learning for Image Anomaly Detection and Segmentation
Chin-Chia Tsai, Tsung-Hsuan Wu, Shang-Hong Lai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3992-4000
Unsupervised representation learning has been proven to be effective for the challenging anomaly detection and segmentation tasks. In this paper, we propose a multi-scale patch-based representation learning method to extract critical and re

presentative information from normal images. By taking the relative feature similarity between patches of different local distances into account, we can achieve better representation learning. Moreover, we propose a refined way to improve the self-supervised learning strategy, thus allowing our model to learn better geometric relationship between neighboring patches. Through sliding patches of different scales all over an image, our model extracts representative features from each patch and compares them with those in the training set of normal images to detect the anomalous regions. Our experimental results on MVTec AD dataset and BTAD dataset demonstrate the proposed method achieves the state-of-the-art accuracy for both anomaly detection and segmentation.

****

Predicting Levels of Household Electricity Consumption in Low-Access Settings

Simone Fobi, Joel Mugyenyi, Nathaniel J. Williams, Vijay Modi, Jay Taneja; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3902-3911

In low-income settings, the most critical piece of information for electric utilities is the anticipated consumption of a customer. Electricity consumption assessment is difficult to do in settings where a significant fraction of households do not yet have an electricity connection. In such settings the absolute levels of anticipated consumption can range from 5-100 kWh/month, leading to high variability amongst these customers. Precious resources are at stake if a significant fraction of low consumers are connected over those with higher consumption. This is the first study of it's kind in low-income settings that attempts to predict a building's consumption and not that of an aggregate administrative area. We train a Convolutional Neural Network (CNN) over pre-electrification daytime satellite imagery with a sample of utility bills from 20,000 geo-referenced electricity customers in Kenya (0.01% of Kenya's residential customers). This is made possible with a two-stage approach that uses a novel building segmentation approach to leverage much larger volumes of no-cost satellite imagery to make the most of scarce and expensive customer data. Our method shows that competitive accuracies can be achieved at the building level, addressing the challenge of consumption variability. This work shows that the building's characteristics and it's surrounding context are both important in predicting consumption levels. We also evaluate the addition of lower resolution geospatial datasets into the training process, including nighttime lights and census-derived data. The results are already helping inform site selection and distribution-level planning, through granular predictions at the level of individual structures in Kenya and there is no reason this cannot be extended to other countries.

****

Neural Radiance Fields Approach to Deep Multi-View Photometric Stereo

Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1965-1977

We present a modern solution to the multi-view photometric stereo problem (MVPS). Our work suitably exploits the image formation model in a MVPS experimental setup to recover the dense 3D reconstruction of an object from images. We procure the surface orientation using a photometric stereo (PS) image formation model and blend it with a multi-view neural radiance field representation to recover the object's surface geometry. Contrary to the previous multi-staged framework to MVPS, where the position, iso-depth contours, or orientation measurements are estimated independently and then fused later, our method is simple to implement and realize. Our method performs neural rendering of multi-view images while utilizing surface normals estimated by a deep photometric stereo network. We render the MVPS images by considering the object's surface normals for each 3D sample point along the viewing direction rather than explicitly using the density gradient in the volume space via 3D occupancy information. We optimize the proposed neural radiance field representation for the MVPS setup efficiently using a fully connected deep network to recover the 3D geometry of an object. Extensive evaluation on the DiLiGenT-MV benchmark dataset shows that our method performs better than the approaches that perform only PS or only multi-view stereo (MVS) and provi des comparable results against the state-of-the-art multi-stage fusion methods.

****

Post-OCR Paragraph Recognition by Graph Convolutional Networks

Renshen Wang, Yasuhisa Fujii, Ashok C. Popat; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 493-502

We propose a new approach for paragraph recognition in document images by spatial graph convolutional networks (GCN) applied on OCR text boxes. Two steps, namely line splitting and line clustering, are performed to extract paragraphs from the lines in OCR results. Each step uses a beta-skeleton graph constructed from bounding boxes, where the graph edges provide efficient support for graph convolution operations. With pure layout input features, the GCN model size is 3 4 orders of magnitude smaller compared to R-CNN based models, while achieving comparable or better accuracies on PubLayNet and other datasets. Furthermore, the GCN models show good generalization from synthetic training data to real-world images, and good adaptivity for variable document styles.

****

PRECODE - A Generic Model Extension To Prevent Deep Gradient Leakage

Daniel Scheliga, Patrick Mäder, Marco Seeland; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1849-1858

Collaborative training of neural networks leverages distributed data by exchanging gradient information between different clients. Although training data entirely resides with the clients, recent work shows that training data can be reconstructed from such exchanged gradient information. To enhance privacy, gradient perturbation techniques have been proposed. However, they come at the cost of reduced model performance, increased convergence time, or increased data demand. In this paper, we introduce PRECODE, a PRivacy EnhanCing mODulE that can be used as generic extension for arbitrary model architectures. We propose a simple yet effective realization of PRECODE using variational modeling. The stochastic sampling induced by variational modeling effectively prevents privacy leakage from gradients and in turn preserves privacy of data owners. We evaluate PRECODE using state of the art gradient inversion attacks on two different model architectures trained on three datasets. In contrast to commonly used defense mechanisms, we find that our proposed modification consistently reduces the attack success rate to 0% while having almost no negative impact on model training and final performance. As a result, PRECODE reveals a promising path towards privacy enhancing model extensions.

****

Evaluation of Correctness in Unsupervised Many-to-Many Image Translation

Dina Bashkirova, Ben Usman, Kate Saenko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1776-1785

Given an input image from a source domain and a guidance image from a target domain, unsupervised many-to-many image-to-image (UMMI2I) translation methods seek to generate a plausible example from the target domain that preserves domain-invariant information of the input source image and inherits the domain-specific information from the guidance image. For example, when translating female faces to male faces, the generated male face should have the same expression, pose and hair color as the input female image, and the same facial hairstyle and other male-specific attributes as the guidance male image. Current state-of-the art UMMI2I methods generate visually pleasing images, but, since for most pairs of real datasets we do not know which attributes are domain-specific and which are domain-invariant, the semantic correctness of existing approaches has not been quantitatively evaluated yet. In this paper, we propose a set of benchmarks and metrics for the evaluation of semantic correctness of these methods. We provide an extensive study of existing state-of-the-art UMMI2I translation methods, showing that all methods, to different degrees, fail to infer which attributes are domain-specific and which are domain-invariant from data, and mostly rely on inductive biases hard-coded into their architectures. Our code can be found at https://github.com/dbash/umi2i_correctness.

****

Discovering Underground Maps From Fashion

Utkarsh Mall, Kavita Bala, Tamara Berg, Kristen Grauman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3114-3123

The fashion sense--meaning the clothing styles people wear--in a geographical region can reveal information about that region. For example, it can reflect the kind of activities people do there, or the type of crowds that frequently visit the region (e.g., tourist hot spot, student neighborhood, business center). We propose a method to create underground neighborhood maps of cities by analyzing how people dress. Using publicly available images from across a city, our method automatically segments the map into neighborhoods with a similar fashion sense. Our approach further allows discovering insights about a city, such as detecting distinct neighborhoods (what is the most unique region of NYC?) and answering analogy questions between cities (what is the "Downtown LA" of Bogota?). We also present two new underground map benchmarks derived from non-image data for 37 cities worldwide. Our method shows promising results on both these benchmarks as well as experiments with human judges.

****

Fast and Efficient Restoration of Extremely Dark Light Fields

Mohit Lamba, Kaushik Mitra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1361-1370

The ability of Light Field (LF) cameras to capture the 3D geometry of a scene in a single photographic exposure has become central to several applications ranging from passive depth estimation to autonomous driving. But these applications cannot rely on LF captured in low-light conditions due to excessive noise and poor image photometry. The existing low-light enhancement techniques are inappropriate for mitigating this problem as they do not leverage LF's multi-view perspective and give blurry restorations. The recent L3Fnet algorithm alleviates this problem reasonably, but its enormous time and memory complexity make it unaffordable for real-world applications. Thus, we propose a three-stage network that is simultaneously much faster and more accurate. We are more accurate because the three stages compute three complementary features: global, local, and view specific features, which are then fused by our RNN inspired feedforward network to restore LF views. We are faster because we restore multiple views simultaneously and so require less number of forward passes. Besides these advantages, our network is flexible enough to restore a m xm LF during inference even if trained for a smaller n xn (n<m) LF without any finetuning. Extensive experiments on real low-light LF demonstrate that compared to state-of-the-art, our model can achieve up to 1 dB higher restoration PSNR, with 9 xspeedup, 23% smaller model size and about 5 xlower floating-point operations.

****

HERS Superpixels: Deep Affinity Learning for Hierarchical Entropy Rate Segmentation

Hankui Peng, Angelica I. Aviles-Rivero, Carola-Bibiane Schönlieb; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 217-226

Superpixels serve as a powerful preprocessing tool in many computer vision tasks. By using superpixel representation, the number of image primitives can be largely reduced by orders of magnitudes. The majority of superpixel methods use handcrafted features, which usually do not translate well into strong adherence to object boundaries. A few recent superpixel methods have introduced deep learning into the superpixel segmentation process. However, none of these methods is able to produce superpixels in near real-time, which is crucial to the applicability of a superpixel method in practice. In this work, we propose a two-stage graph-based framework for superpixel segmentation. In the first stage, we introduce an efficient Deep Affinity Learning (DAL) network that learns pairwise pixel affinities by aggregating multi-scale information. In the second stage, we propose a highly efficient superpixel method called Hierarchical Entropy Rate Segmentation (HERS). Using the learned affinities from the first stage, HERS builds a hierarchical tree structure that can produce any number of highly adaptive superpixels instantaneously. We demonstrate, through visual and numerical experiments, the effectiveness and efficiency of our method compared to various state-of-the-art superpixel methods.

****

Approximate Neural Architecture Search via Operation Distribution Learning

Xingchen Wan, Binxin Ru, Pedro M. Esperança, Fabio Maria Carlucci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2377-2386

The standard paradigm in neural architecture search (NAS) is to search for a fully deterministic architecture with specific operations and connections. In this work, we instead propose to search for the optimal operation distribution, thus providing a stochastic and approximate solution, which can be used to sample architectures of arbitrary length. We propose and show, that given an architectural cell, its performance largely depends on the ratio of used operations, rather than any specific connection pattern; that is, small changes in the ordering of the operations are often irrelevant. This intuition is orthogonal to any specific search strategy and can be applied to a diverse set of NAS algorithms. Through extensive validation on 4 data-sets and 4 NAS techniques (Bayesian optimisation, differentiable search, local search and random search), we show that the operation distribution (1) holds enough discriminating power to reliably identify a solution and (2) is significantly easier to optimise than traditional encodings, leading to large speed-ups at little to no cost in performance. Indeed, this simple intuition significantly reduces the cost of current approaches and potentially enable NAS to be used in a broader range of research applications.

****

Model Compression Using Optimal Transport

Suhas Lohit, Michael Jones; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2764-2773

Model compression methods are important to allow for easier deployment of deep learning models in compute, memory and energy-constrained environments such as mobile phones. Knowledge distillation is a class of model compression algorithms where knowledge from a large teacher network is transferred to a smaller student network thereby improving the student's performance. In this paper, we show how optimal transport-based loss functions can be used for training a student network which encourages learning student network parameters that help bring the distribution of student features closer to that of the teacher features. We present image classification results on CIFAR-100, SVHN and ImageNet and show that the proposed optimal transport loss functions perform comparably to or better than other loss functions.

****

Multi-Dimensional Dynamic Model Compression for Efficient Image Super-Resolution

Zejiang Hou, Sun-Yuan Kung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 633-643

Modern single image super-resolution (SR) system based on convolutional neural networks achieves substantial progress. However, most SR deep networks are computationally expensive and require excessively large activation memory footprints, impeding their effective deployment to resource-limited devices. Based on the observation that the activation patterns in SR networks exhibit high input-dependency, we propose Multi-Dimensional Dynamic Model Compression method that can reduce both spatial and channel wise redundancy in an SR deep network for different input images. To reduce the spatial-wise redundancy, we propose to perform convolution on scaled-down feature-maps where the down-scaling factor is made adaptive to different input images. To reduce the channel-wise redundancy, we introduce a low-cost channel saliency predictor for each convolution to dynamically skip the computation of unimportant channels based on the Gumbel-Softmax. To better capture the feature-maps information and facilitate input-adaptive decision, we employ classic image processing metrics, e.g., Spatial Information, to guide the saliency predictors. The proposed method can be readily applied to a variety of SR deep networks and trained end-to-end with standard super-resolution loss, in combination with a sparsity criterion. Experiments on several benchmarks demonstrate that our method can effectively reduce the FLOPs of both lightweight and no n-compact SR models with negligible PSNR loss. Moreover, our compressed models achieve competitive PSNR-FLOPs Pareto frontier compared with SOTA NAS-based SR methods.

****

Disentangled Representation With Dual-Stage Feature Learning for Face Anti-Spoofing

Yu-Chun Wang, Chien-Yi Wang, Shang-Hong Lai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1955-1964

As face recognition is widely used in diverse security-critical applications, the study of face anti-spoofing (FAS) has attracted more and more attention. Several FAS methods have achieved promising performances if the attack types in the testing data are the same as training data, while the performance significantly degrades for unseen attack types. It is essential to learn more generalized and discriminative features to prevent overfitting to pre-defined spoof attack types. This paper proposes a novel dual-stage disentangled representation learning method that can efficiently untangle spoof-related features from irrelevant ones. Unlike previous FAS disentanglement works with one-stage architecture, we found that the dual-stage training design can improve the training stability and effectively encode the features to detect unseen attack types. Our experiments show that the proposed method provides superior accuracy than the state-of-the-art methods on several cross-type FAS benchmarks.

****

On the Maximum Radius of Polynomial Lens Distortion

Matthew J. Leotta, David Russell, Andrew Matrai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 402-410

Polynomial radial lens distortion models are widely used in image processing and computer vision applications to compensate for when straight lines in the world appear curved in an image. While polynomial models are used pervasively in software ranging from PhotoShop to OpenCV to Blender, they have an often overlooked behavior: polynomial models can fold back onto themselves. This property often goes unnoticed when simply warping to undistort an image. However, in applications such as augmented reality where 3D scene geometry is projected and distorted to overlay an image, this folding can result in a surprising behavior. Points well outside the field of view can project into the middle of the image. The domain of a radial distortion model is only valid up to some (possibly infinite) maximum radius where this folding occurs. This paper derives the closed form expression for the maximum valid radius and demonstrates how this value can be used to filter invalid projections or validate the range of an estimated lens model. Experiments on the popular Lensfun database demonstrate that this folding problem exists on 30% of lens models used in the wild.

****

Generative Adversarial Attack on Ensemble Clustering

Chetan Kumar, Deepak Kumar, Ming Shao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2848-2857

Adversarial attack on learning tasks has attracted substantial attention in recent years; however, most existing works focus on supervised learning. Recently, research has shown that unsupervised learning, such as clustering, tends to be vulnerable due to adversarial attack. In this paper, we focus on a clustering algorithm widely used in the real-world environment, namely, ensemble clustering (EC). EC algorithms usually leverage basic partition (BP) and ensemble techniques to improve the clustering performance collaboratively. Each BP may stem from one trial of clustering, feature segment, or part of data stored on the cloud. We have observed that the attack tends to be less perceivable when only a few BPs are compromised. To explore plausible attack strategies, we propose a novel generative adversarial attack (GA2) model for EC, titled GA2EC. First, we show that not all BPs are equally important, and some of them are more vulnerable under adversarial attack. Second, we develop a generative adversarial model to mimic the attack on EC. In particular, the generative model will simulate behaviors of both clean BPs and perturbed key BPs, and their derived graphs, and thus can launch effective attacks with less attention. We have conducted extensive experiments on eleven clustering benchmarks and have demonstrated that our approach is effective in attacking EC under both transductive and inductive settings.

****

The Hitchhiker's Guide to Prior-Shift Adaptation

Tomáš Šipka, Milan Šulc, Ji■í Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1516-1524

In many computer vision classification tasks, class priors at test time often differ from priors on the training set. In the case of such prior shift, classifiers must be adapted correspondingly to maintain close to optimal performance. This paper analyzes methods for adaptation of probabilistic classifiers to new priors and for estimating new priors on an unlabeled test set. We propose a novel method to address a known issue of prior estimation methods based on confusion matrices, where inconsistent estimates of decision probabilities and confusion matrices lead to negative values in the estimated priors. Experiments on fine-grained image classification datasets provide insight into the best practice of prior shift estimation and classifier adaptation, and show that the proposed method achieves state-of-the-art results in prior adaptation. Applying the best practice to two tasks with naturally imbalanced priors, learning from web-crawled images and plant species classification, increased the recognition accuracy by 1.1% and 3.4% respectively.

****

Hierarchically Decoupled Spatial-Temporal Contrast for Self-Supervised Video Representation Learning

Zehua Zhang, David Crandall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3235-3245

We present a novel technique for self-supervised video representation learning by: (a) decoupling the learning objective into two contrastive subtasks respectively emphasizing spatial and temporal features, and (b) performing it hierarchically to encourage multi-scale understanding. Motivated by their effectiveness in supervised learning, we first introduce spatial-temporal feature learning decoupling and hierarchical learning to the context of unsupervised video learning. We show by experiments that augmentations can be manipulated as regularization to guide the network to learn desired semantics in contrastive learning, and we propose a way for the model to separately capture spatial and temporal features at multiple scales. We also introduce an approach to overcome the problem of divergent levels of instance invariance at different hierarchies by modeling the invariance as loss weights for objective re-weighting. Experiments on downstream action recognition benchmarks on UCF101 and HMDB51 show that our proposed Hierarchically Decoupled Spatial-Temporal Contrast (HDC) makes substantial improvements over directly learning spatial-temporal features as a whole and achieves competitive performance when compared with other state-of-the-art unsupervised methods. Code will be made available.

****

Latent to Latent: A Learned Mapper for Identity Preserving Editing of Multiple Face Attributes in StyleGAN-Generated Images

Siavash Khodadadeh, Shabnam Ghadar, Saeid Motiian, Wei-An Lin, Ladislau Bölöni, Ratheesh Kalarot; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3184-3192

Several recent papers introduced techniques to adjust the attributes of human faces generated by unconditional GANs such as StyleGAN. Despite efforts to disentangle the attributes, a request to change one attribute often triggers unwanted changes to other attributes as well. More importantly, in some cases, a human observer would not recognize the edited face to belong to the same person. We propose an approach where a neural network takes as input the latent encoding of a face and the desired attribute changes and outputs the latent space encoding of the edited image. The network is trained offline using unsupervised data, with training labels generated by an off-the-shelf attribute classifier. The desired attribute changes and conservation laws, such as identity maintenance, are encoded in the training loss. The number of attributes the mapper can simultaneously modify is only limited by the attributes available to the classifier -- we trained

a network that handles 35 attributes, more than any previous approach. As no optimization is performed at deployment time, the computation time is negligible, allowing real-time attribute editing. Qualitative and quantitative comparisons with the current state-of-the-art show our method is better at conserving the identity of the face and restricting changes to the requested attributes.
****

Dynamic Iterative Refinement for Efficient 3D Hand Pose Estimation
John Yang, Yash Bhalgat, Simyung Chang, Fatih Porikli, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1869-1879
While hand pose estimation is a critical component of most interactive extended reality and gesture recognition systems, contemporary approaches are not optimized for computational and memory efficiency. In this paper, we propose a tiny deep neural network of which partial layers are recursively exploited for refining its previous estimations. During its iterative refinements, we employ learned gating criteria to decide whether to exit from the weight-sharing loop, allowing per-sample adaptation in our model. Our network is trained to be aware of the uncertainty in its current predictions to efficiently gate at each iteration, estimating variances after each loop for its keypoint estimates. Additionally, we investigate the effectiveness of end-to-end and progressive training protocols for our recursive structure on maximizing the model capacity. With the proposed setting, our method consistently outperforms state-of-the-art 2D/3D hand pose estimation approaches in terms of both accuracy and efficiency for two widely used benchmarks (e.g., up to 4.9x reduction in GFLOPs and 12.5x fewer parameters than the current SOTA, ACE-Net, while achieving 5.1% AUC improvement on the FPHA dataset).
****

Resource-Efficient Hybrid X-Formers for Vision
Pranav Jeevan, Amit Sethi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2982-2990
Although transformers have become the neural architectures of choice for natural language processing, they require orders of magnitude more training data, GPU memory, and computations in order to compete with convolutional neural networks for computer vision. The attention mechanism of transformers scales quadratically with the length of the input sequence, and unrolled images have long sequence lengths. Plus, transformers lack an inductive bias that is appropriate for images. We tested three modifications to vision transformer (ViT) architectures that address these shortcomings. Firstly, we alleviate the quadratic bottleneck by using linear attention mechanisms, called X-formers (such that, X in Performer, Linformer, Nystromformer), thereby creating Vision X-formers (ViXs). This resulted in up to a seven times reduction in the GPU memory requirement. We also compared their performance with FNet and multi-layer perceptron mixers, which further reduced the GPU memory requirement. Secondly, we introduced an inductive prior for images by replacing the initial linear embedding layer by convolutional layers in ViX, which significantly increased classification accuracy without increasing the model size. Thirdly, we replaced the learnable 1D position embeddings in ViT with Rotary Position Embedding (RoPE), which increases the classification accuracy for the same model size. We believe that incorporating such changes can democratize transformers by making them accessible to those with limited data and computing resources.
****

Quantified Facial Expressiveness for Affective Behavior Analytics
Md Taufeeq Uddin, Shaun Canavan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 985-994
The quantified measurement of facial expressiveness is crucial to analyze human affective behavior at scale. Unfortunately, methods for expressiveness quantification at the video frame-level are largely unexplored, unlike the study of discrete expression. In this work, we propose an algorithm that quantifies facial expressiveness using a bounded, continuous expressiveness score using multimodal facial features, such as action units (AUs), landmarks, head pose, and gaze. The p

roposed algorithm more heavily weights AUs with high intensities and large temporal changes. The proposed algorithm can compute the expressiveness in terms of discrete expression, and can be used to perform tasks including facial behavior tracking and subjectivity quantification in context. Our results on benchmark datasets show the proposed algorithm is effective in terms of capturing temporal changes and expressiveness, measuring subjective differences in context, and extracting useful insight.
****
FalCon: Fine-Grained Feature Map Sparsity Computing With Decomposed Convolutions for Inference Optimization
Zirui Xu, Fuxun Yu, Chenxi Liu, Zhe Wu, Hongcheng Wang, Xiang Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 350-360
Many works focus on the model's static parameter optimization (e.g., filters and weights) for CNN inference acceleration. Compared to parameter sparsity, feature map sparsity is per-input related which has better adaptability. The practical sparsity patterns are non-structural and randomly located on feature maps with non-identical shapes. However, the existing feature map sparsity works take computing efficiency as the primary goal, thereby they can only remove structural sparsity and fail to match the above characteristics. In this paper, we develop a novel sparsity computing scheme called FalCon, which can well adapt to the practical sparsity patterns while still maintaining efficient computing. Specifically, we first propose a decomposed convolution design that enables a fine-grained computing unit for sparsity. Additionally, a decomposed convolution computing optimization paradigm is proposed to convert the sparse computing units to practical acceleration. Extensive experiments show that FalCon achieves at most 67.30% theoretical computation reduction with a neglected accuracy drop while accelerating CNN inference by 37%.
****
METGAN: Generative Tumour Inpainting and Modality Synthesis in Light Sheet Microscopy
Izabela Horvath, Johannes Paetzold, Oliver Schoppe, Rami Al-Maskari, Ivan Ezhov, Suprosanna Shit, Hongwei Li, Ali Ertürk, Bjoern Menze; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 227-237
Novel multimodal imaging methods are capable of generating extensive, super high resolution datasets for preclinical research. Yet, a massive lack of annotations prevents the broad use of deep learning to analyze such data. In this paper, we introduce a novel generative method which leverages real anatomical information to generate realistic image-label pairs of tumours. We construct a dual pathway generator, for the anatomical image and label, trained in a cycle-consistent setup, constrained by an independent, pretrained segmentor. Our method performs two concurrent tasks: domain adaptation and semantic synthesis, which, to our knowledge, has not been done before. The generated images yield significant quantitative improvement compared to existing methods that specialize in either of these tasks. To validate the quality of synthesis, we train segmentation networks on a dataset augmented with the synthetic data, substantially improving the segmentation over the baseline.
****
MoESR: Blind Super-Resolution Using Kernel-Aware Mixture of Experts
Mohammad Emad, Maurice Peemen, Henk Corporaal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3408-3417
Modern deep learning super-resolution approaches have achieved remarkable performance where the low-resolution (LR) input is a degraded high-resolution (HR) image by a fixed known kernel i.e. kernel-specific super-resolution (SR). However, real images often vary in their degradation kernels, thus a single kernel-specific SR approach does not often produce accurate HR results. Recently, degradation-aware networks are introduced to generate blind SR results for unknown kernel conditions. They can restore images for multiple blur kernels, however they have to compromise in quality compared to their kernel-specific counterparts. To addr

ess this issue, we propose a novel blind SR method called Mixture of Experts Super-Resolution (MoESR), which uses different experts for different degradation kernels. A broad space of degradation kernels is covered by kernel-specific SR networks (experts). We present an accurate kernel prediction method (gating mechanism) by evaluating the sharpness of images generated by experts. Based on the predicted kernel our most suited expert network is selected for the input image. Finally, we fine-tune the selected network on the test image itself to leverage the advantage of internal learning. Our experimental results on standard synthetic datasets and real images demonstrate that MoESR outperforms state-of-the-art methods both quantitatively and qualitatively. Especially for the challenging x4 SR task, our PSNR improvement of 0.93 dB on the DIV2KRK dataset is substantial.

****

Spatial-Temporal Transformer for 3D Point Cloud Sequences

Yimin Wei, Hao Liu, Tingting Xie, Qiuhong Ke, Yulan Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1171-1180

Effective learning of spatial-temporal information within a point cloud sequence is highly important for many down-stream tasks such as 4D semantic segmentation and 3D action recognition. In this paper, we propose a novel framework named Point Spatial-Temporal Transformer (PST2) to learn spatial-temporal representations from dynamic 3D point cloud sequences. Our PST2 consists of two major modules: a Spatio-Temporal Self-Attention (STSA) module and a Resolution Embedding (RE) module. Our STSA module is introduced to capture the spatial-temporal context information across adjacent frames, while the RE module is proposed to aggregate features across neighbors to enhance the resolution of feature maps. We test the effectiveness our PST2 with two different tasks on point cloud sequences, i.e., 4D semantic segmentation and 3D action recognition. Extensive experiments on three benchmarks show that our PST2 outperforms existing methods on all datasets. The effectiveness of our STSA and RE modules have also been justified with ablation experiments.

****

LwPosr: Lightweight Efficient Fine Grained Head Pose Estimation

Naina Dhingra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1495-1505

This paper presents a lightweight network for head pose estimation (HPE) task. While previous approaches rely on convolutional neural networks, the proposed network LwPosr uses mixture of depthwise separable convolutional (DSC) and transformer encoder layers which are structured in two streams and three stages to provide fine-grained regression for predicting head poses. The quantitative and qualitative demonstration is provided to show that the proposed network is able to learn head poses efficiently while using less parameter space. Extensive ablations are conducted using three open-source datasets namely 300W-LP, AFLW2000, and BIWI datasets. To our knowledge, (1) LwPosr is the lightest network proposed for estimating head poses compared to both keypoints-based and keypoints-free approaches; (2) it sets a benchmark for both overperforming the previous lightweight network on mean absolute error and on reducing number of parameters; (3) it is first of its kind to use mixture of DSCs and transformer encoders for HPE. This approach is suitable for mobile devices which require lightweight networks.

****

Extractive Knowledge Distillation

Takumi Kobayashi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3511-3520

Knowledge distillation (KD) transfers knowledge of a teacher model to improve performance of a student model which is usually equipped with lower capacity. In the KD framework, however, it is unclear what kind of knowledge is effective and how it is transferred. This paper analyzes a KD process to explore the key factors. In a KD formulation, softmax temperature entangles three main components of student and teacher probabilities and a weight for KD, making it hard to analyze contributions of those factors separately. We disentangle those components so as to further analyze especially the temperature and improve the components respe ctively. Based on the analysis about temperature and uniformity of the teacher probability, we propose a method, called extractive distillation, for extracting effective knowledge from the teacher model. The extractive KD touches only teacher knowledge, thus being applicable to various KD methods. In the experiments on image classification tasks using Cifar-100 and TinyImageNet datasets, we demonstrate that the proposed method outperforms the other KD methods and analyze feature representation to show its effectiveness in the framework of transfer learning.

****

Hessian-Aware Pruning and Optimal Neural Implant

Shixing Yu, Zhewei Yao, Amir Gholami, Zhen Dong, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3880-3891

Pruning is an effective method to reduce the memory footprint and FLOPs associated with neural network models. However, existing structured pruning methods often result in significant accuracy degradation for moderate pruning levels. To address this problem, we introduce a new Hessian Aware Pruning (HAP) method coupled with a Neural Implant approach that uses second-order sensitivity as a metric for structured pruning. The basic idea is to prune insensitive components and to use a Neural Implant for moderately sensitive components, instead of completely pruning them. For the latter approach, the moderately sensitive components are replaced with a low-rank implant that is smaller and less computationally expensive than the original component. We use the relative Hessian trace to measure sensitivity, as opposed to the magnitude-based sensitivity metric commonly used in the literature. We test HAP for both computer vision tasks and natural language tasks, and we achieve new state-of-the-art results. Specifically,HAP achieves less than 0.1%/0.5% degradation on PreResNet29/ResNet50(CIFAR-10/ImageNet) with more than 70%/50% of parameters pruned. Meanwhile, HAP also achieves significantly better performance (up to 0.8% with 60% of parameters pruned) as compared to gradient-based method for head pruning on transformer-based models.

****

PERF-Net: Pose Empowered RGB-Flow Net

Yinxiao Li, Zhichao Lu, Xuehan Xiong, Jonathan Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 513-522

In recent years, many works in the video action recognition literature have shown that two stream models (combining spatial and temporal input streams) are necessary for achieving state-of-the-art performance. In this paper we show the benefits of including yet another stream based on human pose estimated from each frame --- specifically by rendering pose on input RGB frames. At first blush, this additional stream may seem redundant given that human pose is fully determined by RGB pixel values --- however we show (perhaps surprisingly) that this simple and flexible addition can provide complementary gains. Using this insight, we propose a new model, which we dub PERF-Net (short for Pose Empowered RGB-Flow Net), which combines this new pose stream with the standard RGB and flow based input streams via distillation techniques and show that our model outperforms the state-of-the-art by a large margin in a number of human action recognition datasets while not requiring flow or pose to be explicitly computed at inference time. The proposed pose stream is also part of the winner solution of the ActivityNet Kinetics Challenge 2020.

****

Single-Shot Dense Active Stereo With Pixel-Wise Phase Estimation Based on Grid-Structure Using CNN and Correspondence Estimation Using GCN

Ryo Furukawa, Michihiro Mikamo, Ryusuke Sagawa, Hiroshi Kawasaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4001-4011

Active stereo systems based on static pattern projection,a.k.a. oneshot scan, have been widely used for measuring dynamic scenes. Many patterns used for oneshot active stereo have grid structures and grid-wise codes. For such systems, the grid structure is first detected, and graph matching methods are applied to estimate correspondences.However, such graph matching is often vulnerable to graph co

nnection errors caused by grid structure analysis based on image features. Also, dense reconstruction for such systems is an open problem, where pixel-wise correspondence estimation from sparse image features is required. We propose a learning based method to capture grid structure information and pixel-wise positional information simultaneously. We also propose to represent the grid structure by graphs with augmented connections other than 4-neighborconnections and applying them to a graph convolutional network (GCN). Experiments are conducted to confirm the effectiveness of the method by comparing with the existing methods.
****
NUTA: Non-Uniform Temporal Aggregation for Action Recognition
Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Hao Chen, Joseph Tighe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3683-3692
In the world of action recognition research, one primary focus has been on how to construct and train networks to model the spatial-temporal volume of an input video. These methods typically uniformly sample a segment of an input clip (along the temporal dimension). However, not all parts of a video are equally important to determine the action in the clip. In this work, we focus instead on learning where to extract features, so as to focus on the most informative parts of the video. We propose a method called the non-uniform temporal aggregation (NUTA), which aggregates features only from informative temporal segments. We also introduce a synchronization method that allows our NUTA features to be temporally aligned with traditional uniformly sampled video features, so that both local and clip-level features can be combined. Our model has achieved state-of-the-art performance on four widely used large-scale action-recognition datasets (Kinetics400, Kinetics700, Something-something V2 and Charades). In addition, we have created a visualization to illustrate how the proposed NUTA method selects only the most relevant parts of a video clip.
****
Multi-View Fusion of Sensor Data for Improved Perception and Prediction in Autonomous Driving
Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, Carlos Vallespi-Gonzalez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2349-2357
We present an end-to-end method for object detection and trajectory prediction utilizing multi-view representations of LiDAR returns. Our method builds on a state-of-the-art Bird's-Eye View (BEV) network that fuses voxelized features from a sequence of historical LiDAR data as well as rasterized high-definition map to perform detection and prediction tasks. We extend the BEV network with additional LiDAR Range-View (RV) features that use the raw LiDAR information in its native, non-quantized representation. The RV feature map is projected into BEV and fused with the BEV features computed from LiDAR and high-definition map. The fused features are then further processed to output the final detections and trajectories, within a single end-to-end trainable network. In addition, the RV fusion of LiDAR and camera is performed in a straightforward and computational efficient manner using this framework. The proposed approach improves the state-of-the-art on proprietary large-scale real-world data collected by a fleet of self-driving vehicles, as well as on the public nuScenes data set.
****
ADC: Adversarial Attacks Against Object Detection That Evade Context Consistency Checks
Mingjun Yin, Shasha Li, Chengyu Song, M. Salman Asif, Amit K. Roy-Chowdhury, Srikanth V. Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3278-3287
Deep Neural Networks (DNNs) have been shown to be vulnerable to adversarial examples, which are slightly perturbed input images which lead DNNs to make wrong predictions. To protect from such examples, various defense strategies have been proposed. A very recent defense strategy for detecting adversarial examples, that has been shown to be robust to current attacks, is to check for intrinsic context consistencies in the input data, where context refers to various relationship

s (e.g., object-to-object co-occurrence relationships) in images. In this paper, we show that even context consistency checks can be brittle to properly crafted adversarial examples and to the best of our knowledge, we are the first to do so. Specifically, we propose an adaptive framework to generate examples that subvert such defenses, namely, Adversarial attacks against object Detection that evade Context consistency checks (ADC). In ADC, we formulate a joint optimization problem which has two attack goals, viz., (i) fooling the object detector and (ii) evading the context consistency check system, at the same time. Experiments on both PASCAL VOC and MS COCO datasets show that examples generated with ADC fool the object detector with a success rate of over 85% in most cases, and at the same time evade the recently proposed context consistency checks, with a bypassing rate of over 80% in most cases. Our results suggest that how to robustly model context and check its consistency, is still an open problem.

****

Deep Photo Scan: Semi-Supervised Learning for Dealing With the Real-World Degradation in Smartphone Photo Scanning

Man M. Ho, Jinjia Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1880-1889

Physical photographs now can be conveniently scanned by smartphones and stored forever as digital images, yet the scanned photos are not restored well. One solution is to train a supervised deep neural network on many digital images and their smartphone-scanned versions. However, it requires a high labor cost, leading to limited training data. Previous works create training pairs by simulating degradation using low-level image processing techniques. Their synthetic images are then formed with perfectly scanned photos in latent space. Even so, the real-world degradation in smartphone photo scanning remains unsolved since it is more complicated due to lens defocus, low-cost cameras, losing details via printing. Besides, locally structural misalignment still occurs in data due to distorted shapes captured in a 3-D world, reducing restoration performance and the reliability of the quantitative evaluation. To address these problems, we propose a semi-supervised Deep Photo Scan (DPScan). First, we present a way of producing real-world degradation and provide the DIV2K-SCAN dataset for smartphone-scanned photo restoration. Also, Local Alignment is proposed to reduce the minor misalignment remaining in data. Second, we simulate many different variants of the real-world degradation using low-level image transformation to gain a generalization in smartphone-scanned image properties, then train a degradation network to generalize all styles of degradation and provide pseudo-scanned photos for unscanned images as if they were scanned by a smartphone. Finally, we propose a Semi-Supervised Learning that allows our restoration network to be trained on both scanned and unscanned images, diversifying training image content. As a result, the proposed DPScan quantitatively and qualitatively outperforms its baseline architecture, state-of-the-art academic research, and industrial products in smartphone photo scanning.

****

Contextual Gradient Scaling for Few-Shot Learning

Sanghyuk Lee, Seunghyun Lee, Byung Cheol Song; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 834-843

Model-agnostic meta-learning (MAML) is a well-known optimization-based meta-learning algorithm that works well in various computer vision tasks, e.g., few-shot classification. MAML is to learn an initialization so that a model can adapt to a new task in a few steps. However, since the gradient norm of a classifier (head) is much bigger than those of backbone layers, the model focuses on learning the decision boundary of the classifier with similar representations. Furthermore, gradient norms of high-level layers are small than those of the other layers. So, the backbone of MAML usually learns task-generic features, which results in deteriorated adaptation performance in the inner-loop. To resolve or mitigate this problem, we propose contextual gradient scaling (CxGrad), which scales gradient norms of the backbone to facilitate learning task-specific knowledge in the inner-loop. Since the scaling factors are generated from task-conditioned parameters, gradient norms of the backbone can be scaled in a task-wise fashion. Experi mental results show that CxGrad effectively encourages the backbone to learn task-specific knowledge in the inner-loop and improves the performance of MAML up to a significant margin in both same- and cross-domain few-shot classification.

****

MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition

Jiawei Chen, Chiu Man Ho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1910-1921

This paper presents a pure transformer-based approach, dubbed the Multi-Modal Video Transformer (MM-ViT), for video action recognition. Different from other schemes which solely utilize the decoded RGB frames, MM-ViT operates exclusively in the compressed video domain and exploits all readily available modalities, i.e., I-frames, motion vectors, residuals and audio waveform. In order to handle the large number of spatiotemporal tokens extracted from multiple modalities, we develop several scalable model variants which factorize self-attention across the space, time and modality dimensions. In addition, to further explore the rich inter-modal interactions and their effects, we develop and compare three distinct cross-modal attention mechanisms that can be seamlessly integrated into the transformer building block. Extensive experiments on three public action recognition benchmarks (UCF-101,Something-Something-v2, Kinetics-600) demonstrate that MM-ViT outperforms the state-of-the-art video transformers in both efficiency and accuracy, and performs better or equally well to the state-of-the-art CNN counterparts with computationally-heavy optical flow

****

What Makes for Effective Few-Shot Point Cloud Classification?

Chuangguan Ye, Hongyuan Zhu, Yongbin Liao, Yanggang Zhang, Tao Chen, Jiayuan Fan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1829-1838

Due to the emergence of powerful computing resources and large-scale annotated datasets, deep learning has seen wide applications in our daily life. However, most current methods require extensive data collection and retraining when dealing with novel classes never seen before. On the other hand, we humans can quickly recognize new classes by looking at a few samples, which motivates the recent popularity of few-shot learning (FSL) in machine learning communities. Most current FSL approaches work on 2D image domain, however, its implication in 3D perception is relatively under-explored. Not only needs to recognize the unseen examples as in 2D domain, 3D few-shot learning is more challenging with unordered- structures, high intra-class variances and subtle inter-class differences. Moreover, different architectures and learning algorithms make it difficult to study the effectiveness of existing 2D methods when migrating to 3D domain. In this work, for the first time, we perform systematic and extensive studies of recent 2D FSL and 3D backbone networks for benchmarking few-shot point cloud classification, and we suggest a strong baseline and learning architectures for 3D FSL. Then, we propose a novel plug-an and lay component called Cross-Instance Adaptation (CIA) module, to address the subtle inter-class differences and high intra-class variances issues, which can be easily inserted into current baselines with significant performance improvement. Extensive experiments on two newly introduced benchmark datasets, ModelNet40-FS and ShapeNet70-FS, demonstrate the superiority of our proposed network for 3D FSL. Codes and datasets will be released for facilitating future research in this area.

****

edge-SR: Super-Resolution for the Masses

Pablo Navarrete Michelini, Yunhua Lu, Xingqun Jiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1078-1087

Classic image scaling (e.g. bicubic) can be seen as one convolutional layer and a single upscaling filter. Its implementation is ubiquitous in all display devices and image processing software. In the last decade deep learning systems have been introduced for the task of image super-resolution (SR), using several convolutional layers and numerous filters. These methods have taken over the benchmarks of image quality for upscaling tasks. Would it be possible to replace classic upscalers with deep learning architectures on edge devices such as display panels, tablets, laptop computers, etc.? On one hand, the current trend in Edge-AI chips shows a promising future in this direction, with rapid development of hardware that can run deep-learning tasks efficiently. On the other hand, in image SR only few architectures have pushed the limit to extreme small sizes that can actually run on edge devices at real-time. We explore possible solutions to this problem with the aim to fill the gap between classic upscalers and small deep learning configurations. As a transition from classic to deep-learning upscaling we propose edge-SR (eSR), a set of one-layer architectures that use interpretable mechanisms to upscale images. Certainly, a one-layer architecture cannot reach the quality of deep learning systems. Nevertheless, we find that for high speed requirements, eSR becomes better at trading-off image quality and runtime performance. Filling the gap between classic and deep-learning architectures for image upscaling is critical for massive adoption of this technology. It is equally important to have an interpretable system that can reveal the inner strategies to solve this problem and guide us to future improvements and better understanding of larger networks.

****

A Context-Enriched Satellite Imagery Dataset and an Approach for Parking Lot Detection

Yifang Yin, Wenmiao Hu, An Tran, Hannes Kruppa, Roger Zimmermann, See-Kiong Ng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1371-1380

Automatic detection of geoinformation from satellite images has been a fundamental yet challenging problem, which aims to reduce the manual effort of human annotators in maintaining an up-to-date digital map. There are currently several high-resolution satellite imagery datasets that are publicly available. However, the associated ground-truth annotations are limited to road, building, and land use, while the annotations of other geographic objects or attributes are mostly not available. To bridge the gap, we present Grab-Pklot, the first high-resolution and context-enriched satellite imagery dataset for parking lot detection. Our dataset consists of 1344 satellite images with the ground-truth annotations of carparks in Singapore. Motivated by the observation that carparks are mostly co-appear with other geographic objects, we associate each satellite image in our dataset with the surrounding contextual information of road and building, given in the format of multi-channel images. As a side contribution, we present a fusion-based segmentation approach to demonstrate that the parking lot detection accuracy can be improved by modeling the correlations between parking lots and other geographic objects. Experiments on our dataset provide baseline results as well as new insights into the challenges and opportunities in parking lot detection from satellite images.

****

Neural Architecture Search for Efficient Uncalibrated Deep Photometric Stereo

Francesco Sarno, Suryansh Kumar, Berk Kaya, Zhiwu Huang, Vittorio Ferrari, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 361-371

We present an automated machine learning approach for uncalibrated photometric stereo (PS). Our work aims at discovering a light and computationally efficient PS neural network with excellent surface normal accuracy. Unlike previous uncalibrated deep PS networks, which are handcrafted and carefully tuned, we leverage the recent differentiable neural architecture search (NAS) strategy to find uncalibrated PS architecture automatically. We begin by defining a discrete search space for a light calibration network and a normal estimation network, respectively. We then perform a continuous relaxation of this search space, and present a gradient-based optimization strategy to find an efficient light calibration and normal estimation network. Directly applying the NAS methodology to uncalibrated PS is not straightforward as certain mathematical constraints must be satisfied, which we impose explicitly. Moreover, we search for and train the two networks separately to account for the Generalized Bas Relief (GBR) ambiguity. Extensive experiments on the DiLiGenT benchmark show that the automatically searched neura l architectures outperform the current state-of-the-art uncalibrated PS methods, while having a lower memory footprint.

****

AE-StyleGAN: Improved Training of Style-Based Auto-Encoders

Ligong Han, Sri Harsha Musunuri, Martin Renqiang Min, Ruijiang Gao, Yu Tian, Dimitris Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3134-3143

StyleGANs have shown impressive results on data generation and manipulation in recent years, thanks to its disentangled style latent space. A lot of efforts have been made in inverting a pre-trained generator, where an encoder is trained ad hoc after the generator is trained in a two-stage fashion. In this paper, we focus on style-based generators asking a scientific question: Does forcing such a generator to reconstruct real data lead to more disentangled latent space and make the inversion process from image to latent space easy? We describe a new methodology to train a style-based autoencoder where the encoder and generator are optimized end-to-end. We show that our proposed model consistently outperforms baselines in terms of image inversion and generation quality.

****

Agree To Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations

Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 875-884

Deep Learning of neural networks has progressively become more prominent in healthcare with models reaching, or even surpassing, expert accuracy levels. However, these success stories are tainted by concerning reports on the lack of model transparency and bias against some medical conditions or patients' sub-groups. Explainable methods are considered the gateway to alleviate many of these concerns. In this study we demonstrate that the generated explanations are volatile to changes in model training that are perpendicular to the classification task and model structure. This raises further questions about trust in deep learning models for healthcare. Mainly, whether the models capture underlying causal links in the data or just rely on spurious correlations that are made visible via explanation methods. We demonstrate that the output of explainability methods on deep neural networks can vary significantly by changes of hyper-parameters, such as the random seed or how the training set is shuffled. We introduce a measure of explanation consistency which we use to highlight the identified problems on the MIMIC-CXR dataset. We find explanations of identical models but with different training setups have a low consistency: approximately 33% on average. On the contrary, kernel methods are robust against any orthogonal changes, with explanation consistency at 94%. We conclude that current trends in model explanation are not sufficient to mitigate the risks of deploying models in real life healthcare applications.

****

All the Attention You Need: Global-Local, Spatial-Channel Attention for Image Retrieval

Chull Hwan Song, Hye Joo Han, Yannis Avrithis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2754-2763

We address representation learning for large-scale instance-level image retrieval. Apart from backbone, training pipelines and loss functions, popular approaches have focused on different spatial pooling and attention mechanisms, which are at the core of learning a powerful global image representation. There are different forms of attention according to the interaction of elements of the feature tensor (local and global) and the dimensions where it is applied (spatial and channel). Unfortunately, each study addresses only one or two forms of attention and applies it to different problems like classification, detection or retrieval. We present global-local attention module (GLAM), which is attached at the end of a backbone network and incorporates all four forms of attention: local and global, spatial and channel. We obtain a new feature tensor and, by spatial pooling, we learn a powerful embedding for image retrieval. Focusing on global descripto rs, we provide empirical evidence of the interaction of all forms of attention and improve the state of the art on standard benchmarks.

****

REFICS: A Step Towards Linking Vision With Hardware Assurance

Ronald Wilson, Hangwei Lu, Mengdi Zhu, Domenic Forte, Damon L. Woodard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4031-4040

Hardware assurance is a key process in ensuring the integrity, security and functionality of a hardware device. Its heavy reliance on images, especially on Scanning Electron Microscopy images, makes it an excellent candidate for the vision community. The goal of this paper is to provide a pathway for inter-community collaboration by introducing the existing challenges for hardware assurance on integrated circuits in the context of computer vision and support further development using a large-scale dataset with 800,000 images. A detailed benchmark of existing vision approaches in hardware assurance on the dataset is also included for quantitative insights into the problem.

****

Semantically Stealthy Adversarial Attacks Against Segmentation Models

Zhenhua Chen, Chuhua Wang, David Crandall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4080-4089

Segmentation models have been found to be vulnerable to targeted/non-targeted adversarial attacks. However, damaged predictions make it easy to unearth an attack. In this paper, we propose semantically stealthy adversarial attacks which can manipulate targeted labels as designed and preserve non-targeted labels at the same time. In this way, we may hide the corresponding attack behaviors. One challenge is making semantically meaningful manipulations across datasets/models. Another challenge is avoiding damaging non-targeted labels. To solve the above challenges, we consider each input image as prior knowledge to generate perturbations. We also design a special regularizer to help extract features. To evaluate our model's performance, we design three basic attack types, namely `vanishing into the context', `embedding fake labels', and `displacing target objects'. The experiments show that our stealthy adversarial model can attack segmentation models with a relatively high success rate on Cityscapes, Mapillary, and BDD100K. Finally, our framework also shows good generalizations across datasets/models empirically.

****

Adversarial Robustness of Deep Sensor Fusion Models

Shaojie Wang, Tong Wu, Ayan Chakrabarti, Yevgeniy Vorobeychik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2387-2396

We experimentally study the robustness of deep camera-LiDAR fusion architectures for 2D object detection in autonomous driving. First, we find that the fusion model is usually both more accurate, and more robust against single-source attacks than single-sensor deep neural networks. Furthermore, we show that without adversarial training, early fusion is more robust than late fusion, whereas the two perform similarly after adversarial training. However, we note that single-channel adversarial training of deep fusion is often detrimental even to robustness. Moreover, we observe cross-channel externalities, where single-channel adversarial training reduces robustness to attacks on the other channel. Additionally, we observe that the choice of adversarial model in adversarial training is critical: using attacks restricted to cars' bounding boxes is more effective in adversarial training and exhibits less significant cross-channel externalities. Finally, we find that joint-channel adversarial training helps mitigate many of the issues above, but does not significantly boost adversarial robustness.

****

Deep Optimization Prior for THz Model Parameter Estimation

Tak Ming Wong, Hartmut Bauermeister, Matthias Kahl, Peter Haring Bolívar, Michael Möller, Andreas Kolb; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3811-3820

In this paper, we propose a deep optimization prior approach with application to

the estimation of material-related model parameters from terahertz (THz) data that is acquired using a Frequency Modulated Continuous Wave (FMCW) THz scanning system. A stable estimation of the THz model parameters for low SNR and shot noise configurations is essential to achieve acquisition times required for applications in, e.g., quality control. Conceptually, our deep optimization prior approach estimates the desired THz model parameters by optimizing for the weights of a neural network. While such a technique was shown to improve the reconstruction quality for convex objectives in the seminal work of Ulyanov et. al., our paper demonstrates that deep priors also allow to find better local optima in the non-convex energy landscape of the nonlinear inverse problem arising from THz imaging. We verify this claim numerically on various THz parameter estimation problems for synthetic and real data under low SNR and shot noise conditions. While the low SNR scenario not even requires regularization, the impact of shot noise is significantly reduced by total variation (TV) regularization. We compare our approach with existing optimization techniques that require sophisticated physically motivated initialization, and with a 1D single-pixel reparametrization method.
****
Is an Image Worth Five Sentences? A New Look Into Semantics for Image-Text Matching
Ali Furkan Biten, Andrés Mafla, Lluís Gómez, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1391-1400
The task of image-text matching aims to map representations from different modalities into a common joint visual-textual embedding. However, the most widely used datasets for this task, MSCOCO and Flickr30K, are actually image captioning datasets that offer a very limited set of relationships between images and sentences in their ground-truth annotations. This limited ground truth information forces us to use evaluation metrics based on binary relevance: given a sentence query we consider only one image as relevant. However, many other relevant images or captions may be present in the dataset. In this work, we propose two metrics that evaluate the degree of semantic relevance of retrieved items, independently of their annotated binary relevance. Additionally, we incorporate a novel strategy that uses an image captioning metric, CIDEr, to define a Semantic Adaptive Margin (SAM) to be optimized in a standard triplet loss. By incorporating our formulation to existing models, a large improvement is obtained in scenarios where available training data is limited. We also demonstrate that the performance on the annotated image-caption pairs is maintained while improving on other non-annotated relevant items when employing the full training set. The code for our new metric can be found at github.com/furkanbiten/ncs_metric and the model implementation at github.com/andrespmd/semantic_adaptive_margin.
****
Inferring the Class Conditional Response Map for Weakly Supervised Semantic Segmentation
Weixuan Sun, Jing Zhang, Nick Barnes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2878-2887
Image-level weakly supervised semantic segmentation (WSSS) relies on class activation maps (CAMs) for pseudo labels generation. As CAMs only highlight the most discriminative regions of objects, the generated pseudo labels are usually unsatisfactory to serve directly as supervision. To solve this, most existing approaches follow a multi-training pipeline to refine CAMs for better pseudo-labels, which includes: 1) re-training the classification model to generate CAMs; 2) post-processing CAMs to obtain pseudo labels; and 3) training a semantic segmentation model with the obtained pseudo labels. However, this multi-training pipeline requires complicated adjustment and additional time. To address this, we propose a class-conditional inference strategy and an activation aware mask refinement loss function to generate better pseudo labels without re-training the classifier. The class conditional inference-time approach is presented to separately and iteratively reveal the classification network's hidden object activation to generate more complete response maps. Further, our activation aware mask refinement loss function introduces a novel way to exploit saliency maps during segmentation

training and refine the foreground object masks without suppressing background objects. Our method achieves superior WSSS results without requiring re-training of the classifier.

****

Sharing Decoders: Network Fission for Multi-Task Pixel Prediction

Steven Hickson, Karthik Raveendran, Irfan Essa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3771-3780

We examine the benefits of splitting encoder-decoders for multitask learning and showcase results on three tasks (semantics, surface normals, and depth) while adding very few FLOPS per task. Current hard parameter sharing methods for multi-task pixel-wise labeling use one shared encoder with separate decoders for each task. We generalize this notion and term the splitting of encoder-decoder architectures at different points as fission. Our ablation studies on fission show that sharing most of the decoder layers in multi-task encoder-decoder networks results in improvement while adding far fewer parameters per task. Our proposed method trains faster, uses less memory, results in better accuracy, and uses significantly fewer floating point operations (FLOPS) than conventional multi-task methods, with additional tasks only requiring 0.017% more FLOPS than the single-task network. We show results with a real-time model on a Pixel phone with released source code.

****

MovingFashion: A Benchmark for the Video-To-Shop Challenge

Marco Godi, Christian Joppi, Geri Skenderi, Marco Cristani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1678-1686

Retrieving clothes which are worn in social media videos (Instagram, TikTok) is the latest frontier of e-fashion, referred to as "video-to-shop" in the computer vision literature. In this paper we present MovingFashion, the first publicly available dataset to cope with this challenge. MovingFashion is composed of 14855 social videos, each one of them associated to e-commerce "shop" images where the corresponding clothing items are clearly portrayed. In addition, we present a network for retrieving the shop images in this scenario, dubbed SEAM Match-RCNN. The model is trained by image-to-video domain adaptation, allowing to use video sequences where only their association with a shop image is given, eliminating the need of millions of annotated bounding boxes. SEAM Match-RCNN builds an embedding, where an attention-based weighted sum of few frames (10) of a social video is enough to individuate the correct product within the first 5 retrieved items in a 14K+ shop element gallery with an accuracy of 80%. This provides the best performance on MovingFashion, comparing exhaustively against the related state-of-the-art approaches and alternative baselines.

****

VCSeg: Virtual Camera Adaptation for Road Segmentation

Gong Cheng, James H. Elder; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 277-286

Domain shift limits generalization in many problem domains. For road segmentation, one of the principal causes of domain shift is variation in the geometric camera parameters, which results in misregistration of scene structure between images. To address this issue, we decompose the shift into two components: Between-camera shift and within-camera shift. To handle between-camera shift, we assume that average camera parameters are known or can be estimated and use this knowledge to rectify both source and target domain images to a standard virtual camera model. To handle within-camera shift, we use estimates of road vanishing points to correct for shifts in camera pan and tilt. While this approach improves alignment, it produces gaps in the virtual image that complicates network training. To solve this problem, we introduce a novel projective image completion method that fills these gaps in a plausible way. Using five diverse and challenging road segmentation datasets, we demonstrate that our virtual camera method dramatically improves road segmentation performance when generalizing across cameras, and propose that this be integrated as a standard component of road segmentation systems to improve generalization

****

Dual-Head Contrastive Domain Adaptation for Video Action Recognition

Victor G. Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, Elisa Ricci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1181-1190

Unsupervised domain adaptation (UDA) methods have become very popular in computer vision. However, while several techniques have been proposed for images, much less attention has been devoted to videos. This paper introduces a novel UDA approach for action recognition from videos, inspired by recent literature on contrastive learning. In particular, we propose a novel two-headed deep architecture that simultaneously adopts cross-entropy and contrastive losses from different network branches to robustly learn a target classifier. Moreover, this work introduces a novel large-scale UDA dataset, Mixamo->Kinetics, which, to the best of our knowledge, is the first dataset that considers the domain shift arising when transferring knowledge from synthetic to real video sequences. Our extensive experimental evaluation conducted on three publicly available benchmarks and on our new Mixamo->Kinetics dataset demonstrate the effectiveness of our approach, which outperforms the current state-of-the-art methods. Code is available at https://github.com/vturrisi/CO2A.

****

Few-Shot Weakly-Supervised Object Detection via Directional Statistics

Amirreza Shaban, Amir Rahimi, Thalaiyasingam Ajanthan, Byron Boots, Richard Hartley; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3920-3929

Detecting novel objects from few examples has become an emerging topic in computer vision recently. However, current methods need fully annotated training images to learn new object categories which limits their applicability in real world scenarios such as field robotics. In this work, we propose a probabilistic multiple-instance learning approach for few-shot Common Object Localization (COL) and few-shot Weakly Supervised Object Detection (WSOD). In these tasks, only image-level labels, which are much cheaper to acquire, are available. We find that operating on features extracted from the last layer of a pre-trained Faster-RCNN is more effective compared to previous episodic learning based few-shot COL methods. Our model simultaneously learns the distribution of the novel objects and localizes them via expectation-maximization steps. As a probabilistic model, we employ von Mises-Fisher (vMF) distribution which captures the semantic information better than Gaussian distribution when applied to the pre-trained embedding space. When the novel objects are localized, we utilize them to learn a linear appearance model to detect novel classes in new images. Our extensive experiments show that the proposed method, despite being simple, outperforms strong baselines in few-shot COL and WSOD, as well as large-scale WSOD tasks.

****

Pixel-by-Pixel Cross-Domain Alignment for Few-Shot Semantic Segmentation

Antonio Tavera, Fabio Cermelli, Carlo Masone, Barbara Caputo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1626-1635

In this paper we consider the task of semantic segmentation in autonomous driving applications. Specifically, we consider the cross-domain few-shot setting where training can use only few real-world annotated images and many annotated synthetic images. In this context, aligning the domains is made more challenging by the pixel-wise class imbalance that is intrinsic in the segmentation and that leads to ignoring the underrepresented classes and overfitting the well represented ones. We address this problem with a novel framework called Pixel-By-Pixel Cross-Domain Alignment (PixDA). We propose a novel pixel-by-pixel domain adversarial loss following three criteria: (i) align the source and the target domain for each pixel, (ii) avoid negative transfer on the correctly represented pixels, and (iii) regularize the training of infrequent classes to avoid overfitting. The pixel-wise adversarial training is assisted by a novel sample selection procedure, that handles the imbalance between source and target data, and a knowledge distillation strategy, that avoids overfitting towards the few target images. We de

monstrate on standard synthetic-to-real benchmarks that PixDA outperforms previous state-of-the-art methods in (1-5)-shot settings.
****

Lane-Level Street Map Extraction From Aerial Imagery
Songtao He, Hari Balakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2080-2089
Digital maps with lane-level details are the foundation of many applications. However, creating and maintaining digital maps especially maps with lane-level details, are labor-intensive and expensive. In this work, we propose a mapping pipeline to extract lane-level street maps from aerial imagery automatically. Our mapping pipeline first extracts lanes at non-intersection areas, then it enumerates all the possible turning lanes at intersections, validates the connectivity of them, and extracts the valid turning lanes to complete the map. We evaluate the accuracy of our mapping pipeline on a dataset consisting of four U.S. cities, demonstrating the effectiveness of our proposed mapping pipeline and the potential of scalable mapping solutions based on aerial imagery.
****

Let There Be a Clock on the Beach: Reducing Object Hallucination in Image Captioning
Ali Furkan Biten, Lluís Gómez, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1381-1390
Explaining an image with missing or non-existent objects is known as object bias (hallucination) in image captioning. This behaviour is quite common in the state-of-the-art captioning models which is not desirable by humans. To decrease the object hallucination in captioning, we propose three simple yet efficient training augmentation method for sentences which requires no new training data or increase in the model size. By extensive analysis, we show that the proposed methods can significantly diminish our models' object bias on hallucination metrics. Moreover, we experimentally demonstrate that our methods decrease the dependency on the visual features. All of our code, configuration files and model weights is available at https://github.com/furkanbiten/object-bias.
****

Forgery Detection by Internal Positional Learning of Demosaicing Traces
Quentin Bammey, Rafael Grompone von Gioi, Jean-Michel Morel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 328-338
We propose 4Point (Forensics with Positional Internal Training), an unsupervised neural network trained to assess the consistency of the image colour mosaic to find forgeries. Positional learning trains the model to learn the modulo-2 position of pixels, leveraging the translation-invariance of CNN to replicate the underlying mosaic and its potential inconsistencies. Internal learning on a single potentially forged image improves adaption and robustness to varied post-processing and counter-forensics measures. This solution beats existing mosaic detection methods, is more robust to various post-processing and counter-forensic artefacts such as JPEG compression, and can exploit traces to which state-of-the-art generic neural networks are blind. Check qbammey.github.io/4point for the code.
****

Fully Convolutional Cross-Scale-Flows for Image-Based Defect Detection
Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, Bastian Wandt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1088-1097
In industrial manufacturing processes, errors frequently occur at unpredictable times and in unknown manifestations. We tackle this problem, known as automatic defect detection, without requiring any image samples of defective parts. Recent works model the distribution of defect-free image data, using either strong statistical priors or overly simplified data representations. In contrast, our approach handles fine-grained representations incorporating the global and local image context while estimating flexibly the density. To this end, we propose a novel fully convolutional cross-scale normalizing flow (CS-Flow) that jointly proces

ses multiple feature maps of different scales. Using normalizing flows to assign meaningful likelihoods to input samples allows for an efficient defect detection on image-level. Moreover, due to the preserved spatial arrangement the latent space of the normalizing flow is interpretable, i. e. it is applicable to localize defective regions in the image. Our work sets a new state-of-the-art in image-level defect detection on the benchmark datasets Magnetic Tile Defects and MVTec AD showing a 100% AUROC on 4 out of 15 classes.

****

Occlusion-Robust Object Pose Estimation With Holistic Representation
Bo Chen, Tat-Jun Chin, Marius Klimavicius; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2929-2939

Practical object pose estimation demands robustness against occlusions to the target object. State-of-the-art (SOTA) object pose estimators take a two-stage approach, where the first stage predicts 2D landmarks using a deep network and the second stage solves for 6DOF pose from 2D-3D correspondences. Albeit widely adopted, such two-stage approaches could suffer from novel occlusions when generalising and weak landmark coherence due to disrupted features. To address these issues, we develop a novel occlude-and-blackout batch augmentation technique to learn occlusion-robust deep features, and a multi-precision supervision architecture to encourage holistic pose representation learning for accurate and coherent landmark predictions. We perform careful ablation tests to verify the impact of our innovations and compare our method to SOTA pose estimators. Without the need of any post-processing or refinement, our method exhibits superior performance on the LINEMOD dataset. On the YCB-Video dataset our method outperforms all non-refinement methods in terms of the ADD(-S) metric. We also demonstrate the high data-efficiency of our method. Our code is available at http://github.com/BoChenYS/ROPE

****

Fair Visual Recognition in Limited Data Regime Using Self-Supervision and Self-Distillation
Pratik Mazumder, Pravendra Singh, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3095-3103

Deep learning models generally learn the biases present in the training data. Researchers have proposed several approaches to mitigate such biases and make the model fair. Bias mitigation techniques assume that a sufficiently large number of training examples are present. However, we observe that if the training data is limited, then the effectiveness of bias mitigation methods is severely degraded. In this paper, we propose a novel approach to address this problem. Specifically, we adapt self-supervision and self-distillation to reduce the impact of biases on the model in this setting. Self-supervision and self-distillation are not used for bias mitigation. However, through this work, we demonstrate for the first time that these techniques are very effective in bias mitigation. We empirically show that our approach can significantly reduce the biases learned by the model. Further, we experimentally demonstrate that our approach is complementary to other bias mitigation strategies. Our approach significantly improves their performance and further reduces the model biases in the limited data regime. Specifically, on the L-CIFAR-10S skewed dataset, our approach significantly reduces the bias score of the baseline model by 78.22% and outperforms it in terms of accuracy by a significant absolute margin of 8.89%. It also significantly reduces the bias score for the state-of-the-art domain independent bias mitigation method by 59.26% and improves its performance by a significant absolute margin of 7.08%.

****

Robustly Recognizing Irregular Scene Text by Rectifying Principle Irregularities
Changsheng Xu, Yang Wang, Fan Bai, Jihong Guan, Shuigeng Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3061-3068

Reading irregular scene text is a challenging problem in scene text recognition. Rectification is a popular measure to reduce irregularities of text in images.

Existing rectification methods seek to rectify text images into a strictly regular form via free parametric transformation functions. However, they always suffer from information loss or severe deformation due to their poor constraints to the transformation functions. In our investigation, we found that CNN and attention are robust to many slight irregularities. That inspires us to propose a novel and effective rectification method that mainly rectifies the principle regularities, and leaves the slight irregularities to the CNNLSTM-attention recognizer. Our rectification method first estimates the character densities and directions of the input image in a down-sampled map, then finds a best fitting curve from a small predefined Bezier curve set, and finally rectifies the input image with a transformation function corresponding to the selected curve. Transformation functions are carefully designed so that they neither lose important visual information nor cause severe deformation. Extensive experiments on seven benchmark datasets show that our method achieves the state of the art performance in most cases, especially in curved text recognition.
****

Meta Approach to Data Augmentation Optimization
Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2574-2583
Data augmentation policies drastically improve the performance of image recognition tasks, especially when the policies are optimized for the target data and tasks. In this paper, we propose to optimize image recognition models and data augmentation policies simultaneously to improve the performance using gradient descent. Unlike prior methods, our approach avoids using proxy tasks or reducing search space, and can directly improve the validation performance. Our method achieves efficient and scalable training by approximating the gradient of policies by implicit gradient with Neumann series approximation. We demonstrate that our approach can improve the performance of various image classification tasks, including fine-grained image recognition, without using dataset-specific hyperparameter tuning.
****

Fast Nonlinear Image Unblending
Daichi Horita, Kiyoharu Aizawa, Ryohei Suzuki, Taizan Yonetsuji, Huachun Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2051-2059
Nonlinear color blending, which is advanced blending indicated by blend modes such as overlay and multiply, is extensively employed by digital creators to produce attractive visual effects. To enjoy such flexible editing modalities on existing bitmap images like photographs, however, creators need a fast nonlinear blending algorithm that decomposes an image into a set of semi-transparent layers. To address this issue, we propose a neural-network-based method for nonlinear decomposition of an input image into linear and nonlinear alpha layers that can be separately modified for editing purposes, based on the specified color palettes and blend modes. Experiments show that our proposed method achieves an inference speed 370 times faster than the state-of-the-art method of nonlinear image unblending, which uses computationally intensive iterative optimization. Furthermore, our reconstruction quality is higher or comparable than other methods, including linear blending models. In addition, we provide examples that apply our method to image editing with nonlinear blend modes. Our code will be made publicly available.
****

Bayesian Uncertainty and Expected Gradient Length - Regression: Two Sides of the Same Coin?
Megh Shukla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2367-2376
Active learning algorithms select a subset of data for annotation to maximize the model performance on a budget. One such algorithm is Expected Gradient Length, which as the name suggests uses the approximate gradient induced per example in the sampling process. While Expected Gradient Length has been successfully used

for classification and regression, the formulation for regression remains intuitively driven. Hence, our theoretical contribution involves deriving this formulation, thereby supporting experimental evidence [4, 5]. Subsequently, we show that expected gradient length in regression is equivalent to Bayesian uncertainty [22]. If certain assumptions are infeasible, our algorithmic contribution (EGL++) approximates the effect of ensembles with a single deterministic network. Instead of computing multiple possible inferences per input, we leverage previously annotated samples to quantify the probability of previous labels being the true label. Such an approach allows us to extend expected gradient length to a new task: human pose estimation. We perform experimental validation on two human pose datasets (MPII and LSP/LSPET), highlighting the interpretability and competitiveness of EGL++ with different active learning algorithms for human pose estimation.

****

PredStereo: An Accurate Real-Time Stereo Vision System

Diksha Moolchandani, Nivedita Shrivastava, Anshul Kumar, Smruti R. Sarangi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 731-740

Stereo vision algorithms are important building blocks of self-driving applications. The two primary requirements of a self-driving vehicle are real-time operation and nearly 100% accuracy in constructing the 3D scene regardless of the weather conditions and the degree of ambient light. Sadly, most real-time systems as of today provide a level of accuracy that is inadequate and this endangers the life of the passengers; consequently, it is necessary to supplement such systems with expensive LiDAR-based sensors. We observe that for a given scene, different stereo matching algorithms can have vastly different accuracies, and among these algorithms, there is no clear winner. This makes the case for a hybrid stereo vision system where the best stereo vision algorithm for a stereo image pair is chosen by a predictor dynamically, in real-time. We implement such a system called PredStereo in ASIC that combines two diametrically different stereo vision algorithms, CNN-based and traditional, and chooses the best one at runtime. In addition, it associates a confidence with the chosen algorithm, such that the higher-level control system can be switched on in case of a low confidence value. We show that designing a predictor that is explainable and a system that respects soft real-time constraints is non-trivial. Hence, we propose a variety of hardware optimizations that enable our system to work in real-time. Overall, PredStereo improves the disparity estimation error over a state-of-the-art CNN-based stereo vision system by up to 18% (on average 6.25%) with a negligible area overhead (0.003 mm^2) while respecting real-time constraints.

****

Image-Adaptive Hint Generation via Vision Transformer for Outpainting

Daehyeon Kong, Kyeongbo Kong, Kyunghun Kim, Sung-Jun Min, Suk-Ju Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3572-3581

Image outpainting has recently received considerable attention because it can be useful in tasks such as image retargeting and panorama image generation. In general, the problem of extending an image beyond its given boundaries is still ill-posed. Conventional methods predominantly attempt image outpainting by using complex network structures. Some recent studies have tried to decrease the problem complexity through the conversion techniques from outpainting to inpainting. Although these methodologies work well in simple cases, their performance reduces considerably for asymmetrical images. This paper proposes a novel hint-based outpainting methodology that can adaptively select the most plausible patches as hints from a given image to reduce the difficulty of outpainting. To estimate high-quality hints, inspired by patch-based image inpainting methods, we utilize Vision Transformer that also considers self-attention for each patch. The estimated hints are attached on both boundaries of the input image and the inside missing regions are predicted by using an inpainting network. After finishing the prediction, the output image is obtained by removing the hints. Experiments show that our image-adaptive hint framework, when employed in representative inpainting n etworks, can consistently improve its performance compared to the other conversion techniques from outpainting to inpainting on SUN and Beach benchmark datasets.

****

SpectraNet: Learned Recognition of Artificial Satellites From High Contrast Spectroscopic Imagery

J. Zachary Gazak, Ian McQuaid, Ryan Swindle, Matthew Phelps, Justin Fletcher; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4012-4020

Effective space traffic management requires positive identification of artificial satellites. Current methods for extracting object identification from observed data require spatially resolved imagery which limits identification to objects in low earth orbits. Most artificial satellites, however, operate in geostationary orbits at distances which prohibit ground based observatories from resolving spatial information. This paper demonstrates an object identification solution leveraging modified residual convolutional neural networks to map distance-invariant spectroscopic data to object identity. We report classification accuracies exceeding 80% for a simulated 64-class satellite problem--even in the case of satellites undergoing constant, random re-orientation. An astronomical observing campaign driven by these results returned accuracies of 72% for a nine-class problem with an average of 100 examples per class, performing as expected from simulation. We demonstrate the application of variational Bayesian inference by dropout, stochastic weight averaging (SWA), and SWA-focused deep ensembling to measure classification uncertainties--critical components in space traffic management where routine decisions risk expensive space assets and carry geopolitical consequences.

****

S2FGAN: Semantically Aware Interactive Sketch-To-Face Translation

Yan Yang, Md Zakir Hossain, Tom Gedeon, Shafin Rahman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1269-1278

Interactive facial image manipulation attempts to edit single and multiple face attributes using a photo-realistic face and/or semantic mask as input. In the absence of the photo-realistic image (only sketch/mask available), previous methods only retrieve the original face but ignore the potential of aiding model controllability and diversity in the translation process. This paper proposes a sketch-to-image generation framework called S2FGAN, aiming to improve users' ability to interpret and flexibility of face attribute editing from a simple sketch. First, to restore a vivid face from a sketch, we propose semantic level perceptual loss to increase the translation quality. Second, we dedicate the theoretic analysis of attribute editing and build attribute mapping networks with latent semantic loss to modify latent space semantics of Generative Adversarial Networks (GANs). The users can command the model to retouch the generated images by involving the semantic information in the generation process. In this way, our method can manipulate single or multiple face attributes by only specifying attributes to be changed. Extensive experimental results on the CelebAMask-HQ dataset empirically show our superior performance and effectiveness on this task. Our method successfully outperforms state-of-the-art sketch-to-image generation and attribute manipulation methods by exploiting greater control of attribute intensity.

****

Trading-Off Information Modalities in Zero-Shot Classification

Jorge Sánchez, Matías Molina; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3841-3849

Zero-shot classification is the task of learning predictors for classes not seen during training. A practical way to deal with the lack of annotations for the target categories is to encode not only the inputs (images) but also the outputs (object classes) into a suitable representation space. We can use these representations to measure the degree at which images and categories agree by fitting a compatibility measure using the information available during training. One way to define such a measure is by a two step process in which we first project the e lements of either space visual or semantic) onto the other and then compute a similarity score in the target space. Although projections onto the visual space has shown better general performance, little attention has been paid to the degree at which the visual and semantic information contribute to the final predictions. In this paper, we build on this observation and propose two different formulations that allow us to explicitly trade-off the relative importance of the visual and semantic spaces for classification in a zero-shot setting. Our formulations are based on redefinition of the similarity scoring and loss function used to learn the projections. Experiments on six different datasets show that our approach lead to improve performance compared to similar methods. Moreover, combined with synthetic features, our approach competes favorably with the state of the art on both the standard and generalized settings.

****

Co-Segmentation Aided Two-Stream Architecture for Video Captioning

Jayesh Vaidya, Arulkumar Subramaniam, Anurag Mittal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2774-2784

The goal of video captioning is to generate captions for a video by understanding visual and temporal cues. A general video captioning model consists of an Encoder-Decoder framework where Encoder generally captures the visual and temporal information while the decoder generates captions. Recent works have incorporated object-level information into the Encoder by a pretrained off-the-shelf object detector, significantly improving performance. However, using an object detector comes with the following downsides: 1) object detectors may not exhaustively capture all the object categories. 2) In a realistic setting, the performance may be influenced by the domain gap between the object detector and the visual-captioning dataset. To remedy this, we argue that using an external object detector could be eliminated if the model is equipped with the capability of automatically finding salient regions. To achieve this, we propose a novel architecture that learns to attend to salient regions such as objects, persons automatically using a co-segmentation inspired attention module. Then, we utilize a novel salient region interaction module to promote information propagation between salient regions of adjacent frames. Further, we incorporate this salient region-level information into the model using knowledge distillation. We evaluate our model on two benchmark datasets MSR-VTT and MSVD, and show that our model achieves competitive performance without using any object detector.

****

Maximizing Cosine Similarity Between Spatial Features for Unsupervised Domain Adaptation in Semantic Segmentation

Inseop Chung, Daesik Kim, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1351-1360

We propose a novel method that tackles the problem of unsupervised domain adaptation for semantic segmentation by maximizing the cosine similarity between the source and the target domain at the feature level. A segmentation network mainly consists of two parts, a feature extractor and a classification head. We expect that if we can make the two domains have small domain gap at the feature level, they would also have small domain discrepancy at the classification head. Our method computes a cosine similarity matrix between the source feature map and the target feature map, then we maximize the elements exceeding a threshold to guide the target features to have high similarity with the most similar source feature. Moreover, we use a class-wise source feature dictionary which stores the latest features of the source domain to prevent the unmatching problem when computing the cosine similarity matrix and be able to compare a target feature with various source features from various images. Through extensive experiments, we verify that our method gains performance on two unsupervised domain adaptation tasks (GTA5->Cityscaspes and SYNTHIA->Cityscapes).

****

RLSS: A Deep Reinforcement Learning Algorithm for Sequential Scene Generation

Azimkhon Ostonov, Peter Wonka, Dominik L. Michels; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2219-2228

We present RLSS: a reinforcement learning algorithm for sequential scene generation. This is based on employing the proximal policy optimization (PPO) algorithm for generative problems. In particular, we consider how to effectively reduce the action space by including a greedy search algorithm in the learning process. Our experiments demonstrate that our method converges for a relatively large number of actions and learns to generate scenes with predefined design objectives. This approach is placing objects iteratively in the virtual scene. In each step, the network chooses which objects to place and selects positions which result in maximal reward. A high reward is assigned if the last action resulted in desired properties whereas the violation of constraints is penalized. We demonstrate the capability of our method to generate plausible and diverse scenes efficiently by solving indoor planning problems and generating Angry Birds levels.

****

Estimating Image Depth in the Comics Domain

Deblina Bhattacharjee, Martin Everaert, Mathieu Salzmann, Sabine Süsstrunk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2070-2079

Estimating the depth of comics images is challenging as such images a) are monocular; b) lack ground-truth depth annotations; c) differ across different artistic styles; d) are sparse and noisy. We thus, use an off-the-shelf unsupervised image to image translation method to translate the comics images to natural ones and then use an attention-guided monocular depth estimator to predict their depth. This lets us leverage the depth annotations of existing natural images to train the depth estimator. Furthermore, our model learns to distinguish between text and images in the comics panels to reduce text-based artefacts in the depth estimates. Our method consistently outperforms the existing state-of-the-art approaches across all metrics on both the DCM and eBDtheque images. Finally, we introduce a dataset to evaluate depth prediction on comics.

****

GraDual: Graph-Based Dual-Modal Representation for Image-Text Matching

Siqu Long, Soyeon Caren Han, Xiaojun Wan, Josiah Poon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3459-3468

Image-text retrieval task is a challenging task. It aims to measure the visual-semantic correspondence between an image and a text caption. This is tough mainly because the image lacks semantic context information as in its corresponding text caption, and the text representation is very limited to fully describe the details of an image. In this paper, we introduce Graph-based Dual-modal Representations (GraDual), including Vision-Integrated Text Embedding (VITE) and Context-Integrated Visual Embedding (CIVE), for image-text retrieval. The GraDual improves the coverage of each modality by exploiting textual context semantics for the image representation, and using visual features as a guidance for the text representation. To be specific, we design: 1) a dual-modal graph representation mechanism to solve the lack of coverage issue for each modality. 2) an intermediate graph embedding integration strategy to enhance the important pattern across other modality global features. 3) a dual-modal driven cross-modal matching network to generate a filtered representation of another modality. Extensive experiments on two benchmark datasets, MS-COCO and Flickr30K, demonstrates the superiority of the proposed GraDual in comparison to state-of-the-art methods.

****

C-VTON: Context-Driven Image-Based Virtual Try-On Network

Benjamin Fele, Ajda Lampe, Peter Peer, Vitomir Struc; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3144-3153

Image-based virtual try-on techniques have shown great promise for enhancing the user-experience and improving customer satisfaction on fashion-oriented e-commerce platforms. However, they are currently still limited in the quality of the try-on results they are able to produce from input images of diverse characteristics. In this work, we propose a Context-Driven Virtual Try-On Network (C-VTON) that addresses these limitations and convincingly transfers selected clothing ite ms to the target subjects even under challenging pose configurations and in the presence of self-occlusions. At the core of the C-VTON pipeline are: (i) a geometric matching procedure that efficiently aligns the target clothing with the pose of the person in the input images, and (ii) a powerful image generator that utilizes various types of contextual information when synthesizing the final try-on result. C-VTON is evaluated in rigorous experiments on the VITON and MPV datasets and in comparison to state-of-the-art techniques from the literature. Experimental results show that the proposed approach is able to produce photo-realistic and visually convincing results and significantly improves on the existing state-of-the-art.

****

HybVIO: Pushing the Limits of Real-Time Visual-Inertial Odometry

Otto Seiskari, Pekka Rantalankila, Juho Kannala, Jerry Ylilammi, Esa Rahtu, Arno Solin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 701-710

We present HybVIO, a novel hybrid approach for combining filtering-based visual-inertial odometry (VIO) with optimization-based SLAM. The core of our method is highly robust, independent VIO with improved IMU bias modeling, outlier rejection, stationarity detection, and feature track selection, which is adjustable to run on embedded hardware. Long-term consistency is achieved with a loosely-coupled SLAM module. In academic benchmarks, our solution yields excellent performance in all categories, especially in the real-time use case, where we outperform the current state-of-the-art. We also demonstrate the feasibility of VIO for vehicular tracking on consumer-grade hardware using a custom dataset, and show good performance in comparison to current commercial VISLAM alternatives.

****

Semi-Supervised Semantic Segmentation of Vessel Images Using Leaking Perturbations

Jinyong Hou, Xuejie Ding, Jeremiah D. Deng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2625-2634

Semantic segmentation based on deep learning methods can attain appealing accuracy provided large amounts of annotated samples. However, it remains a challenging task when only limited labelled data are available, which is especially common in medical imaging. In this paper, we propose to use Leaking GAN, a GAN-based semi-supervised architecture for retina vessel semantic segmentation. Our key idea is to pollute the discriminator by leaking information from the generator. This leads to more moderate generations that benefit the training of GAN. As a result, the unlabelled examples can be better utilized to boost the learning of the discriminator, which eventually leads to stronger classification performance. In addition, to overcome the variations in medical images, the mean-teacher mechanism is utilized as an auxiliary regularization of the discriminator. Further, we modify the focal loss to fit it as the consistency objective for mean-teacher regularizer. Extensive experiments demonstrate that the Leaking GAN framework achieves competitive performance compared to the state-of-the-art methods when evaluated on benchmark datasets including DRIVE, STARE and CHASE_DB1, using as few as 8 labelled images in the semi-supervised setting. It also outperforms existing algorithms on cross-domain segmentation tasks.

****

SeeTek: Very Large-Scale Open-Set Logo Recognition With Text-Aware Metric Learning

Chenge Li, István Fehérvári, Xiaonan Zhao, Ives Macedo, Srikar Appalaraju; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2544-2553

Recent advances in deep learning and computer vision have set new state of the art in logo recognition. Logo recognition has mostly been approached as a closed-set object recognition problem and more recently as an open-set retrieval problem. Current approaches suffer from distinguishing visually similar logos, especially in open-set retrieval for very large-scale applications with thousands of brands. To address the problem, we propose a multi-task learning architecture of deep metric learning and scene text recognition. We use brand names as weak label s and enforce the model to simultaneously extract distinct visual features as well as predict brand name text. To achieve it, we collected a dataset with 3 Million logos cropped from Amazon Product Catalog images across nearly 8K brands, named PL8K. Our experiments show that adding the task of text recognition during training boosts the model's retrieval performance both on our PL8K dataset and on five other public logo datasets.

****

Measuring Representation of Race, Gender, and Age in Children's Books: Face Detection and Feature Classification in Illustrated Images

Teodora Szasz, Emileigh Harrison, Ping-Jung Liu, Ping-Chang Lin, Hakizumwami Birali Runesha, Anjali Adukia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 462-471

Images in children's books convey messages about society and the roles that people play in it. Understanding these messages requires systematic measurement of who is represented. Computer vision face detection tools can provide such measurements; however, state-of-the-art face detection models were trained with photographs, and 80% of images in children's books are illustrated; thus existing methods both misclassify and miss classifying many faces. In this paper, we introduce a new approach to analyze images using AI tools, resulting in data that can assess representation of race, gender, and age in both illustrations and photographs in children's books. We make four primary contributions to the fields of deep learning and social sciences: (1) We curate an original face detection data set (IllusFace 1.0) by manually labeling 5,403 illustrated faces with bounding boxes. (2) We train two AutoML-based face detection models for illustrations: (i) using IllusFace 1.0 (FDAI); (ii) using iCartoon, a publicly available data set (FDAI_iC), each optimized for illustrated images, detecting 2.5 times more faces in our testing data than the established face detector using Google Vision (FDGV). (3) We curate a data set of the race, gender, and age of 980 faces manually labeled by three different raters (CBFeatures 1.0). (4) We train an AutoML feature classification model (FCA) using CBFeatures 1.0. We compare FCA with the performance of another AutoML model that we trained on UTKFace, a public data set (FCA_UTK) and of an established model using FairFace (FCF). Finally, we examine distributions of character identities over the last century across the models. We find that FCA is 34% more accurate than FCF in its race predictions. These contributions provide tools to educators, caregivers, and curriculum developers to assess the representation contained in children's content.

****

Robust High-Resolution Video Matting With Temporal Guidance

Shanchuan Lin, Linjie Yang, Imran Saleemi, Soumyadip Sengupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 238-247

We introduce a robust, real-time, high-resolution human video matting method that achieves new state-of-the-art performance. Our method is much lighter than previous approaches and can process 4K at 76 FPS and HD at 104 FPS on an Nvidia GTX 1080Ti GPU. Unlike most existing methods that perform video matting frame-by-frame as independent images, our method uses a recurrent architecture to exploit temporal information in videos and achieves significant improvements in temporal coherence and matting quality. Furthermore, we propose a novel training strategy that enforces our network on both matting and segmentation objectives. This significantly improves our model's robustness. Our method does not require any auxiliary inputs such as a trimap or a pre-captured background image, so it can be widely applied to existing human matting applications.

****

Billion-Scale Pretraining With Vision Transformers for Multi-Task Visual Representations

Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, Dmitry Kislyuk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 564-573

Large-scale pretraining of visual representations has led to state-of-the-art performance on a range of benchmark computer vision tasks, yet the benefits of the se techniques at extreme scale in complex production systems has been relatively unexplored. We consider the case of a popular visual discovery product, where these representations are trained with multi-task learning, from use-case specific visual understanding (e.g. skin tone classification) to general representation learning for all visual content (e.g. embeddings for retrieval). In this work, we describe how we (1) generate a dataset with over a billion images via large weakly-supervised pretraining to improve the performance of these visual representations, and (2) leverage Transformers to replace the traditional convolutional backbone, with insights into both system and performance improvements, especially at 1B+ image scale. To support this backbone model, we detail a systematic approach to deriving weakly-supervised image annotations from heterogenous text signals, demonstrating the benefits of clustering techniques to handle the long-tail distribution of image labels. Through a comprehensive study of offline and online evaluation, we show that large-scale Transformer-based pretraining provides significant benefits to industry computer vision applications. The model is deployed in a production visual shopping system, with 36% improvement in top-1 relevance and 23% improvement in click-through volume. We conduct extensive experiments to better understand the empirical relationships between Transformer-based architectures, dataset scale, and the performance of production vision systems.

****

Towards Active Vision for Action Localization With Reactive Control and Predictive Learning

Shubham Trehan, Sathyanarayanan N. Aakur; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 783-792

Visual event perception tasks such as action localization have primarily focused on supervised learning settings under a static observer, i.e., the camera is static and cannot be controlled by an algorithm. They are often restricted by the quality, quantity, and diversity of annotated training data and do not often generalize to out-of-domain samples. In this work, we tackle the problem of active action localization where the goal is to localize an action while controlling the geometric and physical parameters of an active camera to keep the action in the field of view without training data. We formulate an energy-based mechanism that combines predictive learning and reactive control to perform active action localization without rewards, which can be sparse or non-existent in real-world environments. We perform extensive experiments in both simulated and real-world environments on two tasks - active object tracking and active action localization. We demonstrate that the proposed approach can generalize to different tasks and environments in a streaming fashion, requiring only a single pass through the video, working in real-time. We show that the proposed approach outperforms unsupervised baselines and obtains competitive performance compared to those trained with reinforcement learning.

****

Learnable Multi-Level Frequency Decomposition and Hierarchical Attention Mechanism for Generalized Face Presentation Attack Detection

Meiling Fang, Naser Damer, Florian Kirchbuchner, Arjan Kuijper; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3722-3731

With the increased deployment of face recognition systems in our daily lives, face presentation attack detection (PAD) is attracting much attention and playing a key role in securing face recognition systems. Despite the great performance achieved by the hand-crafted and deep-learning-based methods in intra-dataset evaluations, the performance drops when dealing with unseen scenarios. In this work, we propose a dual-stream convolution neural networks (CNNs) framework. One stream adapts four learnable frequency filters to learn features in the frequency domain, which are less influenced by variations in sensors/illuminations. The other stream leverages the RGB images to complement the features of the frequency domain. Moreover, we propose a hierarchical attention module integration to join the information from the two streams at different stages by considering the nature of deep features in different layers of the CNN. The proposed method is evaluated in the intra-dataset and cross-dataset setups, and the results demonstrate that our proposed approach enhances the generalizability in most experimental setups in comparison to state-of-the-art, including the methods designed explicitly for domain adaption/shift problems. We successfully prove the design of our proposed PAD solution in a step-wise ablation study that involves our proposed learnable frequency decomposition, our hierarchical attention module design, and the used loss function. Training codes and pre-trained models are publicly released.

****

Skeleton-DML: Deep Metric Learning for Skeleton-Based One-Shot Action Recognition

Raphael Memmesheimer, Simon Häring, Nick Theisen, Dietrich Paulus; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3702-3710

One-shot action recognition allows the recognition of human-performed actions with only a single training example. This can influence human-robot-interaction positively by enabling the robot to react to previously unseen behavior. We formulate the one-shot action recognition problem as a deep metric learning problem and propose a novel image-based skeleton representation that performs well in a metric learning setting. Therefore, we train a model that projects the image representations into an em-bedding space. In embedding space, similar actions have a low euclidean distance while dissimilar actions have a higher distance. The one-shot action recognition problem becomes a nearest-neighbor search in a set of activity reference samples. We evaluate the performance of our pro-posed representation against a variety of other skeleton-based image representations. In addition, we present an ablation study that shows the influence of different embedding vector sizes, losses and augmentation. Our approach lifts the state-of-the-art by 3.3% for the one-shot action recognition protocol on the NTU RGB+D 120 dataset under a comparable training setup. With additional augmentation, our result improved over 7.7%

****

Adversarial Semantic Hallucination for Domain Generalized Semantic Segmentation

Gabriel Tjio, Ping Liu, Joey Tianyi Zhou, Rick Siow Mong Goh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 318-327

Convolutional neural networks typically perform poorly when the test (target domain) and training (source domain) data have significantly different distributions. While this problem can be mitigated by using the target domain data to align the source and target domain feature representations, the target domain data may be unavailable due to privacy concerns. Consequently, there is a need for methods that generalize well despite restricted access to target domain data during training. In this work, we propose an adversarial semantic hallucination approach (ASH), which combines a class-conditioned hallucination module and a semantic segmentation module. Since the segmentation performance varies across different classes, we design a semantic-conditioned style hallucination module to generate affine transformation parameters from semantic information in the segmentation probability maps of the source domain image. Unlike previous adaptation approaches, which treat all classes equally, ASH considers the class-wise differences. The segmentation module and the hallucination module compete adversarially, with the hallucination module generating increasingly "difficult" stylized images to challenge the segmentation module. In response, the segmentation module improves as it is trained with generated samples at an appropriate class-wise difficulty level. Our results on the Cityscapes and Mapillary benchmark datasets show that our method is competitive with state of the art work. Code is made available at https://github.com/gabriel-tjio/ASH.

****

Self-Supervised Generative Style Transfer for One-Shot Medical Image Segmentation

Devavrat Tomar, Behzad Bozorgtabar, Manana Lortkipanidze, Guillaume Vray, Mohammad Saeed Rad, Jean-Philippe Thiran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1998-2008

In medical image segmentation, supervised deep networks' success comes at the cost of requiring abundant labeled data. While asking domain experts to annotate only one or a few of the cohort's images is feasible, annotating all available images is impractical. This issue is further exacerbated when pre-trained deep networks are exposed to a new image dataset from an unfamiliar distribution. Using available open-source data for ad-hoc transfer learning or hand-tuned techniques for data augmentation only provides suboptimal solutions. Motivated by atlas-based segmentation, we propose a novel volumetric self-supervised learning for data augmentation capable of synthesizing volumetric image-segmentation pairs via learning transformations from a single labeled atlas to the unlabeled data. Our work's central tenet benefits from a combined view of one-shot generative learning and the proposed self-supervised training strategy that cluster unlabeled volumetric images with similar styles together. Unlike previous methods, our method does not require input volumes at inference time to synthesize new images. Instead, it can generate diversified volumetric image-segmentation pairs from a prior distribution given a single or multi-site dataset. Augmented data generated by our method used to train the segmentation network provide significant improvements over state-of-the-art deep one-shot learning methods on the task of brain MRI segmentation. Ablation studies further exemplified that the proposed appearance model and joint training are crucial to synthesize realistic examples compared to existing medical registration methods. The code, data, and models are available at https://github.com/devavratTomar/SST/.
****

Semi-Supervised Domain Adaptation via Sample-to-Sample Self-Distillation
Jeongbeen Yoon, Dahyun Kang, Minsu Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1978-1987

Semi-supervised domain adaptation (SSDA) is to adapt a learner to a new domain with only a small set of labeled samples when a large labeled dataset is given on a source domain. In this paper, we propose a pair-based SSDA method that adapts a model to the target domain using self-distillation with sample pairs. Each sample pair is composed of a teacher sample from a labeled dataset (i.e., source or labeled target) and its student sample from an unlabeled dataset (i.e., unlabeled target). Our method generates an assistant feature by transferring an intermediate style between the teacher and the student, and then train the model by minimizing the output discrepancy between the student and the assistant. During training, the assistants gradually bridge the discrepancy between the two domains, thus allowing the student to easily learn from the teacher. Experimental evaluation on standard benchmarks shows that our method effectively minimizes both the inter-domain and intra-domain discrepancies, thus achieving significant improvements over recent methods.
****

MobileStereoNet: Towards Lightweight Deep Networks for Stereo Matching
Faranak Shamsafar, Samuel Woerz, Rafia Rahim, Andreas Zell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2417-2426

Recent methods in stereo matching have continuously improved the accuracy using deep models. This gain, however, is attained with a high increase in computation cost, such that the network may not fit even on a moderate GPU. This issue raises problems when the model needs to be deployed on resource-limited devices. For this, we propose two light models for stereo vision with reduced complexity and without sacrificing accuracy. Depending on the dimension of cost volume, we design a 2D and a 3D model with encoder-decoders built from 2D and 3D convolutions, respectively. To this end, we leverage 2D MobileNet blocks and extend them to 3D for stereo vision application. Besides, a new cost volume is proposed to boost the accuracy of the 2D model, making it performing close to 3D networks. Experiments show that the proposed 2D/3D networks effectively reduce the computational expense (27%/95% and 72%/38% fewer parameters/operations in 2D and 3D models, respectively) while upholding the accuracy. Code: https://github.com/cogsys-tuebingen/mobilestereonet.
****

In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation
Jan Weyler, Federico Magistri, Peter Seitz, Jens Behley, Cyrill Stachniss; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2725-2734

A detailed analysis of a plant's phenotype in real field conditions is critical for plant scientists and breeders to understand plant function. In contrast to traditional phenotyping performed manually, vision-based systems have the potential for an objective and automated assessment with high spatial and temporal resolution. One of such systems' objectives is to detect and segment individual leaves of each plant since this information correlates to the growth stage and provides phenotypic traits, such as leaf count, coverage, and size. In this paper, we propose a vision-based approach that performs instance segmentation of individual crop leaves and associates each with its corresponding crop plant in real fields. This enables us to compute relevant basic phenotypic traits on a per-plant level. We employ a convolutional neural network and operate directly on drone imagery. The network generates two different representations of the input image that we utilize to cluster individual crop leaf and plant instances. We propose a novel method to compute clustering regions based on our network's predictions that achieves high accuracy. Furthermore, we compare to other state-of-the-art approaches and show that our system achieves superior performance. The source code of our approach is available.

****

Visually Guided Sound Source Separation and Localization Using Self-Supervised Motion Representations
Lingyu Zhu, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1289-1299

In this paper, we perform audio-visual sound source separation, i.e. to separate component audios from a mixture based on the videos of sound sources. Moreover, we aim to pinpoint the source location in the input video sequence. Recent works have shown impressive audio-visual separation results when using prior knowledge of the source type (e.g. human playing instrument) and pre-trained motion detectors (e.g. keypoints or optical flows). However, at the same time, the models are limited to a certain application domain. In this paper, we address these limitations and make the following contributions: i) we propose a two-stage architecture, called Appearance and Motion network (AMnet), where the stages specialise to appearance and motion cues, respectively. The entire system is trained in a self-supervised manner; ii) we introduce an Audio-Motion Embedding (AME) framework to explicitly represent the motions that related to sound; iii) we propose an audio-motion transformer architecture for audio and motion feature fusion; iv) we demonstrate state-of-the-art performance on two challenging datasets (MUSIC-21 and AVE) despite the fact that we do not use any pre-trained keypoint detectors or optical flow estimators. Project page: https://ly-zhu.github.io/self-supervised-motion-representations

****

BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection
Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, Jongwon Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 48-57

The advancement in numerous generative models has a two-fold effect: a simple and easy generation of realistic synthesized images, but also an increased risk of malicious abuse of those images. Thus, it is important to develop a generalized detector for synthesized images of any GAN model or object category, including those unseen during the training phase. However, the conventional methods heavily depend on the training settings, which cause a dramatic decline in performance when tested with unknown domains. To resolve the issue and obtain a generalized detection ability, we propose Bilateral High-Pass Filters (BiHPF), which amplify the effect of the frequency-level artifacts that are generally found in the synthesized images of generative models. Also, to find the properties of the general frequency-level artifacts, we develop an additional method to adversarially extract the artifact compression map. Numerous experimental results validate that our method outperforms other state-of-the-art methods, even when tested with unseen domains.

****

Weakly-Supervised Convolutional Neural Networks for Vessel Segmentation in Cerebral Angiography

Arvind Vepa, Andrew Choi, Noor Nakhaei, Wonjun Lee, Noah Stier, Andrew Vu, Greyson Jenkins, Xiaoyan Yang, Manjot Shergill, Moira Desphy, Kevin Delao, Mia Levy, Cristopher Garduno, Lacy Nelson, Wandi Liu, Fan Hung, Fabien Scalzo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 585-594

Automated vessel segmentation in cerebral digital subtraction angiography (DSA) has significant clinical utility in the management of cerebrovascular diseases such as stroke diagnosis and detection of aneurysms. While deep learning is state-of-the-art in segmentation, a significant amount of labeled data is needed for training. Because of domain differences, pretrained networks cannot be applied to DSA data out-of-the-box. We propose a novel learning framework, which utilizes an active contour model for weak supervision and low-cost human-in-the-loop strategies to improve weak label quality. Our study produces several significant results, including state-of-the-art results for cerebral DSA vessel segmentation, which exceed human annotator quality, and an analysis of annotation cost and model performance trade-offs utilizing weak supervision strategies. Additionally, we will be publicly releasing code to reproduce our methodology and our dataset, the largest known high-quality annotated cerebral DSA vessel segmentation dataset.

****

Towards a Robust Differentiable Architecture Search Under Label Noise

Christian Simon, Piotr Koniusz, Lars Petersson, Yan Han, Mehrtash Harandi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3256-3266

We all have experienced the difficulty of designing appropriate neural architectures due to the lack of general principles and best practices. The game changer might be touse Neural Architecture Search (NAS) where a machine does all the hard work for us based on the data at its disposal. Invarious problems and in particular in classification, architectures designed by NAS outperform or compete with the best manual network designs in terms of accuracy, size, memory footprint and FLOPs. That said, previous studies focus ondeveloping NAS algorithms for clean high quality data, a restrictive and somewhat unrealistic assumption. In this paper, focusing on the differentiable NAS algorithms, we show that vanilla NAS algorithms suffer from a performance loss if class labels are noisy. To combat this issue, we propose tomake use of the principle of information bottleneck as a regularizer. This leads us to develop a noise injecting operation that is included during the learning process, preventing the network from learning from noisy samples. Our empirical evaluations show that the noise injecting operation does not degrade the performance of the NAS algorithm if the data is indeed clean. In contrast, if the data is noisy, the architecture learned by our algorithm comfortably outperforms algorithms specifically equipped with sophisticated mechanisms to learn in the presence of label noise. In contrast to many algorithms designed to work in the presence of noisylabels, prior knowledge about the properties of the noise and its characteristics are not required for our algorithm.

****

AFTer-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation

Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3971-3981

Recent advances in transformer-based models have drawn attention to exploring these techniques in medical image segmentation, especially in conjunction with the U-Net model (or its variants), which has shown great success in medical image segmentation, under both 2D and 3D settings. Current 2D based methods either directly replace convolutional layers with pure transformers or consider a transformer as an additional intermediate encoder between the encoder and decoder of U-Ne

t. However, these approaches only consider the attention encoding within one single slice and do not utilize the axial-axis information naturally provided by a 3D volume. In the 3D setting, convolution on volumetric data and transformers both consume large GPU memory. One has to either downsample the image or use cropped local patches to reduce GPU memory usage, which limits its performance. In this paper, we propose Axial Fusion Transformer UNet (AFTer-UNet), which takes both advantages of convolutional layers' capability of extracting detailed features and transformers' strength on long sequence modeling. It considers both intra-slice and inter-slice long-range cues to guide the segmentation. Meanwhile, it has fewer parameters and takes less GPU memory to train than the previous transformer-based models. Extensive experiments on three multi-organ segmentation datasets demonstrate that our method outperforms current state-of-the-art methods.
****
Variational Stacked Local Attention Networks for Diverse Video Captioning
Tonmoay Deb, Akib Sadmanee, Kishor Kumar Bhaumik, Amin Ahsan Ali, M Ashraful Amin, A K M Mahbubur Rahman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4070-4079
While describing spatio-temporal events in natural language, video captioning models mostly rely on the encoder's latent visual representation. Recent progress on the encoder-decoder model attends encoder features mainly in linear interaction with the decoder. However, growing model complexity for visual data encourages more explicit feature interaction for fine-grained information, which is currently absent in the video captioning domain. Moreover, feature aggregations methods have been used to unveil richer visual representation, either by the concatenation or using a linear layer. Though feature sets for a video semantically overlap to some extent, these approaches result in objective mismatch and feature redundancy. In addition, diversity in captions is a fundamental component of expressing one event from several meaningful perspectives, currently missing in the temporal, i.e., video captioning domain. To this end, we propose Variational Stacked Local Attention Network (VSLAN), which exploits low-rank bilinear pooling for self-attentive feature interaction and stacking multiple video feature streams in a discount fashion. Each feature stack's learned attributes contribute to our proposed diversity encoding module, followed by the decoding query stage to facilitate end-to-end diverse and natural captions without any explicit supervision on attributes. We evaluate VSLAN on MSVD and MSR-VTT datasets in terms of syntax and diversity. The CIDEr score of VSLAN outperforms current off-the-shelf methods by 7.8% on MSVD and 4.5% on MSR-VTT, respectively. On the same datasets, VSLAN achieves competitive results in caption diversity metrics.
****
A Modular and Unified Framework for Detecting and Localizing Video Anomalies
Keval Doshi, Yasin Yilmaz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3982-3991
Anomaly detection in videos has been attracting an increasing amount of attention. Despite the competitive performance of recent methods on benchmark datasets, they typically lack desirable features such as modularity, cross-domain adaptivity, interpretability, and real-time anomalous event detection. Furthermore, current state-of-the-art approaches are evaluated using the standard instance-based detection metric by considering video frames as independent instances, which is not ideal for video anomaly detection. Motivated by these research gaps, we propose a modular and unified approach to the online video anomaly detection and localization problem, called MOVAD, which consists of a novel transfer learning based plug-and-play architecture, a sequential anomaly detector, a mathematical framework for selecting the detection threshold, and a suitable performance metric for real-time anomalous event detection in videos. Extensive performance evaluations on benchmark datasets show that the proposed framework significantly outperforms the current state-of-the-art approaches.
****
FastAno: Fast Anomaly Detection via Spatio-Temporal Patch Transformation
Chaewon Park, MyeongAh Cho, Minhyeok Lee, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2249-

2259

Video anomaly detection has gained significant attention due to the increasing requirements of automatic monitoring for surveillance videos. Especially, the prediction based approach is one of the most studied methods to detect anomalies by predicting frames that include abnormal events in the test set after learning with the normal frames of the training set. However, a lot of prediction networks are computationally expensive owing to the use of pre-trained optical flow networks, or fail to detect abnormal situations because of their strong generative ability to predict even the anomalies. To address these shortcomings, we propose spatial rotation transformation (SRT) and temporal mixing transformation (TMT) to generate irregular patch cuboids within normal frame cuboids in order to enhance the learning of normal features. Additionally, the proposed patch transformation is used only during the training phase, allowing our model to detect abnormal frames at fast speed during inference. Our model is evaluated on three anomaly detection benchmarks, achieving competitive accuracy and surpassing all the previous works in terms of speed.

****

Low-Cost Multispectral Scene Analysis With Modality Distillation

Heng Zhang, Elisa Fromont, Sébastien Lefèvre, Bruno Avignon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 803-812

Despite its robust performance under various illumination conditions, multispectral scene analysis has not been widely deployed due to two strong practical limitations: 1) thermal cameras, especially high-resolution ones are much more expensive than conventional visible cameras; 2) the most commonly adopted multispectral architectures, two-stream neural networks, nearly double the inference time of a regular mono-spectral model which makes them impractical in embedded environments. In this work, we aim to tackle these two limitations by proposing a novel knowledge distillation framework named Modality Distillation (MD). The proposed framework distils the knowledge from a high thermal resolution two-stream network with feature-level fusion to a low thermal resolution one-stream network with image-level fusion. We show on different multispectral scene analysis benchmarks that our method can effectively allow the use of low-resolution thermal sensors with more compact one-stream networks.

****

Single-Shot Path Integrated Panoptic Segmentation

Sukjun Hwang, Seoung Wug Oh, Seon Joo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3328-3337

Panoptic segmentation, which is a novel task of unifying instance segmentation and semantic segmentation, has attracted a lot of attention lately. However, most of the previous methods are composed of multiple pathways with each pathway specialized to a designated segmentation task. In this paper, we propose to resolve panoptic segmentation in single-shot by integrating the execution flows. With the integrated pathway, a unified feature map called Panoptic-Feature is generated, which includes the information of both things and stuffs. Panoptic-Feature becomes more sophisticated by auxiliary problems that guide to cluster pixels that belong to the same instance and differentiate between objects of different classes. A collection of convolutional filters, where each filter represents either a thing or stuff, is applied to Panoptic-Feature at once, materializing the single-shot panoptic segmentation. Taking the advantages of both top-down and bottom-up approaches, our method, named SPINet, enjoys high efficiency and accuracy on major panoptic segmentation benchmarks: COCO and Cityscapes.

****

Knowledge Capture and Replay for Continual Learning

Saisubramaniam Gopalakrishnan, Pranshu Ranjan Singh, Haytham Fayek, Savitha Ramasamy, ArulMurugan Ambikapathi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 10-18

Deep neural networks model data for a task or a sequence of tasks, where the knowledge extracted from the data is encoded in the parameters and representations of the network. Extraction and utilization of these representations is vital whe n data is no longer available in the future, especially in a continual learning scenario. We introduce 'flashcards', which are visual representations that 'capture' the encoded knowledge of a network as a recursive function of some predefined random image patterns. In a continual learning scenario, flashcards help to prevent catastrophic forgetting by consolidating the knowledge of all the previous tasks. Flashcards are required to be constructed only before learning the subsequent task, hence, they are independent of the number of tasks trained before, making them task agnostic. We demonstrate the efficacy of flashcards in capturing learned knowledge representation (as an alternative to the original data), and empirically validate on a variety of continual learning tasks: reconstruction, denoising, and task-incremental classification, using several heterogeneous (varying background and complexity) benchmark datasets. Experimental evidence indicates that: (i) flashcards as a replay strategy is 'task agnostic', (ii) performs better than generative replay, and (iii) is on par with episodic replay without additional memory overhead.

****

3DRefTransformer: Fine-Grained Object Identification in Real-World Scenes Using Natural Language

Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3941-3950

In this paper, we study fine-grained 3D object identification in real-world scenes described by a textual query. The task aims to discriminatively understand an instance of a particular 3D object described by natural language utterances among other instances of 3D objects of the same class appearing in a visual scene. We introduce the 3DRefTransformer net, a transformer-based neural network that identifies 3D objects described by linguistic utterances in real-world scenes. The network's input is 3D object segmented point cloud images representing a real-world scene and a language utterance that refers to one of the scene objects. The goal is to identify the referred object. Compared to the state-of-the-art models that are mostly based on graph convolutions and LSTMs, our 3DRefTransformer net offers two key advantages. First, it is an end-to-end transformer model that operates both on language and 3D visual objects. Second, it has a natural ability to ground textual terms in the utterance to the learning representation of 3D objects in the scene. We further incorporate object pairwise spatial relation loss and contrastive learning during model training. We show in our experiments that our model improves the performance upon the current SOTA significantly on Referit3D Nr3D and Sr3D datasets. Code and Models will be made publicly available.

****

Self-Supervised Pretraining Improves Self-Supervised Pretraining

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, Kurt Keutzer, Trevor Darrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2584-2594

While self-supervised pretraining has proven beneficial for many computer vision tasks, it requires expensive and lengthy computation, large amounts of data, and is sensitive to data augmentation. Prior work demonstrates that models pretrained on datasets dissimilar to their target data, such as chest X-ray models trained on ImageNet, underperform models trained from scratch. Users that lack the resources to pretrain must use existing models with lower performance. This paper explores Hierarchical PreTraining (HPT), which decreases convergence time and improves accuracy by initializing the pretraining process with an existing pretrained model. Through experimentation on 16 diverse vision datasets, we show HPT converges up to 80x faster, improves accuracy across tasks, and improves the robustness of the self-supervised pretraining process to changes in the image augmentation policy or amount of pretraining data. Taken together, HPT provides a simple framework for obtaining better pretrained representations with less computational resources.

****

Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using

Counterfactuals

Saloni Dash, Vineeth N Balasubramanian, Amit Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 915-924

Counterfactual examples for an input---perturbations that change specific features but not others---have been shown to be useful for evaluating bias of machine learning models, e.g., against specific demographic groups. However, generating counterfactual examples for images is non-trivial due to the underlying causal structure on the various features of an image. To be meaningful, generated perturbations need to satisfy constraints implied by the causal model. We present a method for generating counterfactuals by incorporating a structural causal model (SCM) in an improved variant of Adversarially Learned Inference (ALI), that generates counterfactuals in accordance with the causal relationships between attributes of an image. Based on the generated counterfactuals, we show how to explain a pre-trained machine learning classifier, evaluate its bias, and mitigate the bias using a counterfactual regularizer. On the Morpho-MNIST dataset, our method generates counterfactuals comparable in quality to prior work on SCM-based counterfactuals. Our method also works on the more complex CelebA faces dataset. Generated counterfactuals are indistinguishable from reconstructed images in a human evaluation experiment and we use them to evaluate a standard classifier trained on CelebA data. We show that the classifier is biased w.r.t. skin and hair color, and how counterfactual regularization can remove those biases.

****

PROVES: Establishing Image Provenance Using Semantic Signatures

Mingyang Xie, Manav Kulshrestha, Shaojie Wang, Jinghan Yang, Ayan Chakrabarti, Ning Zhang, Yevgeniy Vorobeychik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 543-552

Modern AI tools, such as generative adversarial networks, have transformed our ability to create and modify visual data with photorealistic results. However, one of the deleterious side-effects of these advances is the emergence of nefarious uses in manipulating information in visual data, such as through the use of deep fakes. We propose a novel architecture for preserving the provenance of semantic information in images to make them less susceptible to deep fake attacks. Our architecture includes semantic signing and verification steps. We apply this architecture to verifying two types of semantic information: individual identities (faces) and whether the photo was taken indoors or outdoors. Verification in both cases carefully accounts for a collection of common image transformation, such as translation, scaling, cropping, and small rotations, and rejects adversarial transformations, such as adversarially perturbed or, in the case of face verification, swapped faces. Experiments demonstrate that in the case of provenance of faces in an image, our approach is robust to black-box adversarial transformations (which are rejected) as well as benign transformations (which are accepted), with few false negatives and false positives. Background verification, on the other hand, is susceptible to black-box adversarial examples, but becomes significantly more robust after adversarial training.

****

Class-Balanced Active Learning for Image Classification

Javad Zolfaghari Bengar, Joost van de Weijer, Laura Lopez Fuentes, Bogdan Raducanu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1536-1545

Active learning aims to reduce the labeling effort that is required to train algorithms by learning an acquisition function selecting the most relevant data for which a label should be requested from a large unlabeled data pool. Active learning is generally studied on balanced datasets where an equal amount of images per class is available. However, real-world datasets suffer from severe imbalanced classes, the so called long-tail distribution. We argue that this further complicates the active learning process, since the imbalanced data pool can result in suboptimal classifiers. To address this problem in the context of active learning, we proposed a general optimization framework that explicitly takes class-balancing into account. Results on three datasets showed that the method is general (it can be combined with most existing active learning algorithms) and can be effectively applied to boost the performance of both informative and representative-based active learning methods. In addition, we showed that also on balanced datasets our method generally results in a performance gain.

****

Tailor Me: An Editing Network for Fashion Attribute Shape Manipulation

Youngjoong Kwon, Stefano Petrangeli, Dahun Kim, Haoliang Wang, Viswanathan Swaminathan, Henry Fuchs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3831-3840

Fashion attribute editing aims to manipulate fashion images based on a user-specified attribute, while preserving the details of the original image as intact as possible. Recent works in this domain have mainly focused on direct manipulation of the raw RGB pixels, which only allows to perform edits involving relatively small shape changes (e.g., sleeves). The goal of our Virtual Personal Tailoring Network (VPTNet) is to extend the editing capabilities to much larger shape changes of fashion items, such as cloth length. To achieve this goal, we decouple the fashion attribute editing task into two conditional stages: shape-then-appearance editing. To this aim, we propose a shape editing network that employs a semantic parsing of the fashion image as an interface for manipulation. Compared to operating on the raw RGB image, our parsing map editing enables performing more complex shape editing operations. Second, we introduce an appearance completion network that takes the previous stage results and completes the shape difference regions to produce the final RGB image. Qualitative and quantitative experiments on the DeepFashion-Synthesis dataset confirm that VPTNet outperforms state-of-the-art methods for both small and large shape attribute editing.

****

Addressing Out-of-Distribution Label Noise in Webly-Labelled Data

Paul Albert, Diego Ortego, Eric Arazo, Noel E. O'Connor, Kevin McGuinness; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 392-401

A recurring focus of the deep learning community is towards reducing the labeling effort. Data gathering and annotation using a search engine is a simple alternative to generating a fully human-annotated and human-gathered dataset. Although web crawling is very time efficient, some of the retrieved images are unavoidably noisy, i.e. incorrectly labeled. Designing robust algorithms for training on noisy data gathered from the web is an important research perspective that would render the building of datasets easier. In this paper we conduct a study to understand the type of label noise to expect when building a dataset using a search engine. We review the current limitations of state-of-the-art methods for dealing with noisy labels for image classification tasks in the case of web noise distribution. We propose a simple solution to bridge the gap with a fully clean dataset using Dynamic Softening of Out-of-distribution Samples (DSOS), which we design on corrupted versions of the CIFAR-100 dataset, and compare against state-of-the-art algorithms on the web noise perturbated MiniImageNet and Stanford datasets and on real label noise datasets: WebVision 1.0 and Clothing1M. Our work is fully reproducible https://git.io/JKGcj.

****

Learning Maritime Obstacle Detection From Weak Annotations by Scaffolding

Lojze Žust, Matej Kristan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 955-964

Coastal water autonomous boats rely on robust perception methods for obstacle detection and timely collision avoidance. The current state-of-the-art is based on deep segmentation networks trained on large datasets. Per-pixel ground truth labeling of such datasets, however, is labor-intensive and expensive. We observe that far less information is required for practical obstacle avoidance -- the location of water edge on static obstacles like shore and approximate location and bounds of dynamic obstacles in the water is sufficient to plan a reaction. We propose a new scaffolding learning regime (SLR) that allows training obstacle detection segmentation networks only from such weak annotations, thus significantly reducing the cost of ground-truth labeling. Experiments show that maritime obstacle segmentation networks trained using SLR substantially outperform the same ne tworks trained with dense ground truth labels, despite a significant reduction in labelling effort. Thus accuracy is not sacrificed for labelling simplicity but is in fact improved, which is a remarkable result.

****

Revealing Disocclusions in Temporal View Synthesis Through Infilling Vector Prediction

Vijayalakshmi Kanchana, Nagabhushan Somraj, Suraj Yadwad, Rajiv Soundararajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3541-3550

We consider the problem of temporal view synthesis, where the goal is to predict a future video frame from the past frames using knowledge of the depth and relative camera motion. In contrast to revealing the disoccluded regions through intensity based infilling, we study the idea of an infilling vector to infill by pointing to a non-disoccluded region in the synthesized view. To exploit the structure of disocclusions created by camera motion during their infilling, we rely on two important cues, temporal correlation of infilling directions and depth. We design a learning framework to predict the infilling vector by computing a temporal prior that reflects past infilling directions and a normalized depth map as input to the network. We conduct extensive experiments on a large scale dataset we build for evaluating temporal view synthesis in addition to the SceneNet RGB-D dataset. Our experiments demonstrate that our infilling vector prediction approach achieves superior quantitative and qualitative infilling performance compared to other approaches in literature. Our dataset and code can be found at https://nagabhushansn95.github.io/publications/2021/ivp.html

****

Generating and Controlling Diversity in Image Search

Md. Mehrab Tanjim, Ritwik Sinha, Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W. Cottrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 411-419

In our society, generations of systemic biases have led to some professions being more common among certain genders and races. This bias is also reflected in image search on stock image repositories and search engines, e.g., a query like "male Asian administrative assistant" may produce limited results. The pursuit of a utopian world demands providing content users with an opportunity to present any profession with diverse racial and gender characteristics. The limited choice of existing content for certain combinations of profession, race, and gender presents a challenge to content providers. Current research dealing with bias in search mostly focuses on re-ranking algorithms. However, these methods cannot create new content or change the overall distribution of protected attributes in photos. To remedy these problems, we propose a new task of high-fidelity image generation by controlling multiple attributes from imbalanced datasets. Our proposed task poses new sets of challenges for the state-of-the-art Generative Adversarial Networks (GANs). In this paper, we also propose a new training framework to better address the challenges. We evaluate our framework rigorously on a real-world dataset and perform user studies that show our model is preferable to the alternatives.

****

Visualizing Paired Image Similarity in Transformer Networks

Samuel Black, Abby Stylianou, Robert Pless, Richard Souvenir; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3164-3173

Transformer architectures have shown promise for a wide range of computer vision tasks, including image embedding. As was the case with convolutional neural networks and other models, explainability of the predictions is a key concern, but visualization approaches tend to be architecture-specific. In this paper, we introduce a new method for producing interpretable visualizations that, given a pair of images encoded with a Transformer, show which regions contributed to their similarity. Additionally, for the task of image retrieval, we compare the performance of Transformer and ResNet models of similar capacity and show that while they have similar performance in aggregate, the retrieved results and the visual explanations for those results are quite different. Code is available at https://github.com/vidarlab/xformer-paired-viz.

****

Knowledge-Augmented Contrastive Learning for Abnormality Classification and Localization in Chest X-Rays With Radiomics Using a Feedback Loop

Yan Han, Chongyan Chen, Ahmed Tewfik, Benjamin Glicksberg, Ying Ding, Yifan Peng, Zhangyang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2465-2474

Accurate classification and localization of abnormalities in chest X-rays play an important role in clinical diagnosis and treatment planning. Building a highly accurate predictive model for these tasks usually requires a large number of manually annotated labels and pixel regions (bounding boxes) of abnormalities. However, it is expensive to acquire such annotations, especially the bounding boxes. Recently, contrastive learning has shown strong promise in leveraging unlabeled natural images to produce highly generalizable and discriminative features. However, extending its power to the medical image domain is under-explored and highly non-trivial, since medical images are much less amendable to data augmentations. In contrast, their prior knowledge, as well as radiomic features, is often crucial. To bridge this gap, we propose an end-to-end semi-supervised knowledge-augmented contrastive learning framework, that simultaneously performs disease classification and localization tasks. The key knob of our framework is a unique positive sampling approach tailored for the medical images, by seamlessly integrating radiomic features as a knowledge augmentation. Specifically, we first apply an image encoder to classify the chest X-rays and to generate the image features. We next leverage Grad-CAM to highlight the crucial (abnormal) regions for chest X-rays (even when unannotated), from which we extract radiomic features. The radiomic features are then passed through another dedicated encoder to act as the positive sample for the image features generated from the same chest X-ray. In this way, our framework constitutes a feedback loop for image and radiomic features to mutually reinforce each other. Their contrasting yields knowledge-augmented representations that are both robust and interpretable. Extensive experiments on the NIH Chest X-ray dataset demonstrate that our approach outperforms existing baselines in both classification and localization tasks.

****

A Structure-Aware Method for Direct Pose Estimation

Hunter Blanton, Scott Workman, Nathan Jacobs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2019-2028

Estimating camera pose from a single image is a fundamental problem in computer vision. Existing methods for solving this task fall into two distinct categories, which we refer to as direct and indirect. Direct methods, such as PoseNet, regress pose from the image as a fixed function, for example using a feed-forward convolutional network. Such methods are desirable because they are deterministic and run in constant time. Indirect methods for pose regression are often non-deterministic, with various external dependencies such as image retrieval and hypothesis sampling. We propose a direct method that takes inspiration from structure-based approaches to incorporate explicit 3D constraints into the network. Our approach maintains the desirable qualities of other direct methods while achieving much lower error in general. Code is available https://github.com/mvrl/structure-aware-pose-estimation.

****

Late-Resizing: A Simple but Effective Sketch Extraction Strategy for Improving Generalization of Line-Art Colorization

Dohyun Kim, Dajung Je, Kwangjin Lee, Moohyun Kim, Han Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3469-3478

Automatic line-art colorization is a demanding research field owing to its expensive and labor-intensive workload. Learning-based approaches have lately emerged to improve the quality of colorization. To handle the lack of paired data in line art and color images, sketch extraction has been widely adopted. This study primarily focuses on the resizing process applied within the sketch extraction pr ocedure, which is essential for normalizing input sketches of various sizes to the target size of the colorization model. We first analyze the inherent risk in a conventional resizing strategy, i.e., early-resizing, which places the resizing step before the line detection process to ensure the practicality. Although the strategy is extensively used, it involves an often overlooked risk of significantly degrading the generalization of the colorization model. Thus, we propose a late-resizing strategy in which resizing is applied after the line detection step. The proposed late-resizing strategy has three advantages: prevention of a quality degradation in the color image, augmentation for downsizing artifacts, and alleviation of look-ahead bias. In conclusion, we present both quantitative and qualitative evaluations on representative learning-based line-art colorization methods, which verify the effectiveness of the proposed method in the generalization of the colorization model.

****

Temporally Stable Video Segmentation Without Video Annotations

Aharon Azulay, Tavi Halperin, Orestis Vantzos, Nadav Bornstein, Ofir Bibi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3449-3458

Temporally consistent dense video annotations are scarce and hard to collect. In contrast, image segmentation datasets (and pre-trained models) are ubiquitous, and easier to label for any novel task. In this paper, we introduce a method to adapt still image segmentation models to video in an unsupervised manner, by using an optical flow-based consistency measure. To ensure that the inferred segmented videos appear more stable in practice, we verify that the consistency measure is well correlated with human judgement via a user study. Training a new multi-input multi-output decoder using this measure as a loss, together with a technique for refining current image segmentation datasets and a temporal weighted-guided filter, we observe stability improvements in the generated segmented videos with minimal loss of accuracy.

****

Fair and Accurate Age Prediction Using Distribution Aware Data Curation and Augmentation

Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, Yuval Elovici; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3551-3561

Deep learning-based facial recognition systems have experienced increased media attention due to exhibiting unfair behavior. Large enterprises, such as IBM, shut down their facial recognition and age prediction systems as a consequence. Age prediction is an especially difficult application with the issue of fairness remaining an open research problem (e.g. predicting age for different ethnicity equally accurate). One of the main causes of unfair behavior in age prediction methods lies in the distribution and diversity of the training data. In this work, we present two novel approaches for dataset curation and data augmentation in order to increase fairness through balanced feature curation and increase diversity through distribution aware augmentation. To achieve this, we introduce out-of-distribution detection to the facial recognition domain which is used to select the data most relevant to the deep neural network's (DNN) task when balancing the data among age, ethnicity, and gender. Our approach shows promising results. Our best-trained DNN model outperformed all academic and industrial baselines in terms of fairness by up to 4.92 times and also enhanced the DNN's ability to generalise outperforming Amazon AWS and Microsoft Azure public cloud systems by 31.88% and 10.95%, respectively.

****

Re-Compose the Image by Evaluating the Crop on More Than Just a Score

Yang Cheng, Qian Lin, Jan P. Allebach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1-9

Image re-composition has always been regarded as one of the most important steps during the post-processing of a photo. The quality of an image re-composition mainly depends on a person's taste in aesthetics, which is not an effortless task for those who have no abundant experience in photography. Besides, while re-com posing one image does not require much of a person's time, it could be quite time-consuming when there are hundreds of images to be re-composed. To solve these problems, we propose a method that automates the process of re-composing an image to the desired aspect ratio. Although there already exist many image re-composition methods, they only provide a score to their predicted best crop but fail to explain why the score is high or low. Conversely, we succeed in designing an explainable method by introducing a novel 10-layer aesthetic score map, which represents how the position of the saliency in the original uncropped image, relative to that of the crop region, contributes to the overall score of the crop, so that the crop is not just represented by a single score. We conducted experiments to show that the proposed score map boosts the performance of our algorithm, which achieves a state-of-the-art performance on both public and our own datasets.

****

AttWalk: Attentive Cross-Walks for Deep Mesh Analysis

Ran Ben Izhak, Alon Lahav, Ayellet Tal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1546-1555

Mesh representation by random walks has been shown to benefit deep learning. Randomness is indeed a powerful concept. However, it comes with a price--some walks might wander around non-characteristic regions of the mesh, which might be harmful to shape analysis, especially when only a few walks are utilized. We propose a novel walk-attention mechanism that leverages the fact that multiple walks are used for a single mesh representation. The key idea is that the walks may provide each other with information regarding the meaningful (attentive) features of the mesh. We utilize this mutual information to extract a single descriptor of the mesh. This differs from common attention mechanisms that use attention to improve the representation of each individual descriptor. Our approach achieves SoTA results for two basic 3D shape analysis tasks: classification and retrieval. Even a handful of walks along a mesh suffice for learning. Furthermore, our approach provides insight into mesh importance detection.

****

Extraction of Positional Player Data From Broadcast Soccer Videos

Jonas Theiner, Wolfgang Gritz, Eric Müller-Budack, Robert Rein, Daniel Memmert, Ralph Ewerth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 823-833

Computer-aided support and analysis are becoming increasingly important in the modern world of sports. The scouting of potential prospective players, performance as well as match analysis, and the monitoring of training programs rely more and more on data-driven technologies to ensure success. Therefore, many approaches require large amounts of data, which are, however, not easy to obtain in general. In this paper, we propose a pipeline for the fully-automated extraction of positional data from broadcast video recordings of soccer matches. In contrast to previous work, the system integrates all necessary sub-tasks like sports field registration, player detection, or team assignment that are crucial for player position estimation. The quality of the modules and the entire system is interdependent. A comprehensive experimental evaluation is presented for the individual modules as well as the entire pipeline to identify the influence of errors to subsequent modules and the overall result. In this context, we propose novel evaluation metrics to compare the output with ground-truth positional data.

****

Self-Guidance: Improve Deep Neural Network Generalization via Knowledge Distillation

Zhenzhu Zheng, Xi Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3203-3212

We present Self-Guidance, a simple way to train deep neural networks via knowledge distillation. The basic idea is to train sub-network to match the prediction of the full network, so-called "Self-Guidance". Under the "teacher-student" framework, we construct both teacher and student within the same target network. Student network is the sub-networks that randomly skip some portions of the full network. The teacher network is the full network, can be considered as the ensembl

e of all possible student networks. The training process is performed in a closed-loop: (1) Forward prediction contains two passes that generate student and teacher predictions. (2) Backward distillation allows knowledge transfer from the teacher back to students. Comprehensive evaluations show that our approach improves the generalization ability of deep neural networks to a significant margin. The results prove our superior performance in both image classification on CIFAR10, CIFAR100, and facial expression recognition on FER-2013 and RAF.
****
mToFNet: Object Anti-Spoofing With Mobile Time-of-Flight Data
Yonghyun Jeong, Doyeon Kim, Jaehyeon Lee, Minki Hong, Solbi Hwang, Jongwon Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 38-47
In online markets, sellers can maliciously recapture others' images on display screens to utilize as spoof images, which can be challenging to distinguish in human eyes. To prevent such harm, we propose an anti-spoofing method using the pairs of RGB images and depth maps provided by the mobile camera with a time-of-fight sensor. When images are recaptured on display screens, various patterns differing by the screens as known as the moire patterns can be also captured in spoof images. These patterns lead the anti-spoofing model to be overfitted and unable to detect spoof images recaptured on unseen media. To avoid the issue, we build a novel representation model composed of two embedding models, which can be trained without considering the recaptured images. Also, we newly introduce mToF dataset, the largest and most diverse object anti-spoofing dataset, and the first to utilize the time-of-flight (ToF) data. Experimental results confirm that our model achieves robust generalization even across unseen domains.
****
Towards Durability Estimation of Bioprosthetic Heart Valves via Motion Symmetry Analysis
Maryam Alizadeh, Melissa Cote, Alexandra Branzan Albu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3124-3133
This paper addresses bioprosthetic heart valve (BHV) durability estimation via computer vision (CV)-based analyses of the visual symmetry of valve leaflet motion. BHVs are routinely implanted in patients suffering from valvular heart diseases. Valve designs are rigorously tested using cardiovascular equipment, but once implanted, more than 50% of BHVs encounter a structural failure within 15 years. We investigate the correlation between the visual dynamic symmetry of BHV leaflets and the functional symmetry of the valves. We hypothesize that an asymmetry in the valve leaflet motion will generate an asymmetry in the flow patterns, resulting in added local stress and forces on some of the leaflets, which can accelerate the failure of the valve. We propose two different pair-wise leaflet symmetry scores based on the diagonals of orthogonal projection matrices (DOPM) and on dynamic time warping (DTW), computed from videos recorded during pulsatile flow tests. We compare the symmetry score profiles with those of fluid dynamic parameters (velocity and vorticity values) at the leaflet borders, obtained from valve-specific numerical simulations. Experiments on four cases that include three different tricuspid BHVs yielded promising results, with the DTW scores showing a good coherence with respect to the simulations. With a link between visual and functional symmetries established, this approach paves the way towards BHV durability estimation using CV techniques.
****
ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection
Danila Rukhovich, Anna Vorontsova, Anton Konushin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2397-2406
In this paper, we introduce the task of multi-view RGB-based 3D object detection as an end-to-end optimization problem. To address this problem, we propose ImVoxelNet, a novel fully convolutional method of 3D object detection based on posed monocular or multi-view RGB images. The number of monocular images in each multi-view input can variate during training and inference; actually, this number mi

ght be unique for each multi-view input. ImVoxelNet successfully handles both indoor and outdoor scenes, which makes it general-purpose. Specifically, it achieves state-of-the-art results in car detection on KITTI (monocular) and nuScenes (multi-view) benchmarks among all methods that accept RGB images. Moreover, it surpasses existing RGB-based 3D object detection methods on the SUN RGB-D dataset. On ScanNet, ImVoxelNet sets a new benchmark for multi-view 3D object detection.

****

Time-Space Transformers for Video Panoptic Segmentation

Andra Petrovai, Sergiu Nedevschi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 925-934

We propose a novel solution for the task of video panoptic segmentation, that simultaneously predicts pixel-level semantic and instance segmentation and generates clip-level instance tracks. Our network, named VPS-Transformer, with a hybrid architecture based on the state-of-the-art panoptic segmentation network Panoptic-DeepLab, combines a convolutional architecture for single-frame panoptic segmentation and a novel video module based on an instantiation of the pure Transformer block. The Transformer, equipped with attention mechanisms, models spatio-temporal relations between backbone output features of current and past frames for more accurate and consistent panoptic estimates. As the pure Transformer block introduces large computation overhead when processing high resolution images, we propose a few design changes for a more efficient compute. We study how to aggregate information more effectively over the space-time volume and we compare several variants of the Transformer block with different attention schemes. Extensive experiments on the Cityscapes-VPS dataset demonstrate that our best model improves the temporal consistency and video panoptic quality by a margin of 2.2%, with little extra computation.

****

Pro-CCaps: Progressively Teaching Colourisation to Capsules

Rita Pucci, Christian Micheloni, Gian Luca Foresti, Niki Martinel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2271-2279

Automatic image colourisation studies how to colourise greyscale images. Existing approaches exploit convolutional layers that extract image-level features learning the colourisation on the entire image, but miss entities-level ones due to pooling strategies. We believe that entity-level features are of paramount importance to deal with the intrinsic multimodality of the problem (i.e., the same object can have different colours, and the same colour can have different properties). Models based on capsule layers aim to identify entity-level features in the image from different points of view, but they do not keep track of global features. Our network architecture integrates entity-level features into the image-level features to generate a plausible image colourisation. We observed that results obtained with direct integration of such two representations are largely dominated by the image-level features, thus resulting in unsaturated colours for the entities. To limit such an issue, we propose a gradual growth of the reconstruction phase of the model while training. By advantaging of prior knowledge from each growing step, we obtain a stable collaboration between image-level and entity-level features that ultimately generates stable and vibrant colourisations. Experimental results on three benchmark datasets, and a user study, demonstrate that our approach has competitive performance with respect to the state-of-the-art and provides more consistent colourisation. Code available at omitted-for-reviewing-purposes.

****

Resolution-Robust Large Mask Inpainting With Fourier Convolutions

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, Victor Lempitsky; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2149-2159

Modern image inpainting systems, despite the significant progress, often struggle with large missing areas, complex geometric structures, and high-resolution images. We find that one of the main reasons for that is the lack of an effective receptive field in both the inpainting network and the loss function. To alleviate this issue, we propose a new method called large mask inpainting (LaMa). LaMa is based on i) a new inpainting network architecture that uses fast Fourier convolutions (FFCs), which have the image-wide receptive field; ii) a high receptive field perceptual loss; iii) large training masks, which unlocks the potential of the first two components. Our inpainting network improves the state-of-the-art across a range of datasets and achieves excellent performance even in challenging scenarios, e.g. completion of periodic structures. Our model generalizes surprisingly well to resolutions that are higher than those seen at train time, and achieves this at lower parameter & time costs than the competitive baselines. The code is available at https://github.com/saic-mdal/lama.

****

Calibrating CNNs for Few-Shot Meta Learning
Peng Yang, Shaogang Ren, Yang Zhao, Ping Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2090-2099

Although few-shot meta learning has been extensively studied in machine learning community, the fast adaptation towards new tasks remains a challenge in the few-shot learning scenario. The neuroscience research reveals that the capability of evolving neural network formulation is essential for task adaptation, which has been broadly studied in recent meta-learning researches. In this paper, we present a novel forward-backward meta-learning framework (FBM) to facilitate the model generalization in few-shot learning from a new perspective, i.e., neuron calibration. In particular, FBM models the neurons in deep neural network-based model as calibrated units under a general formulation, where neuron calibration could empower fast adaptation capability to the neural network-based models through influencing both their forward inference path and backward propagation path. The proposed calibration scheme is lightweight and applicable to various feed-forward neural network architectures. Extensive empirical experiments on the challenging few-shot learning benchmarks validate that our approach training with neuron calibration achieves a promising performance, which demonstrates that neuron calibration plays a vital role in improving the few-shot learning performance.

****

Fast and Explicit Neural View Synthesis
Pengsheng Guo, Miguel Angel Bautista, Alex Colburn, Liang Yang, Daniel Ulbricht, Joshua M. Susskind, Qi Shan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3791-3800

We study the problem of novel view synthesis from sparse source observations of a scene comprised of 3D objects. We propose a simple yet effective approach that is neither continuous nor implicit, challenging recent trends on view synthesis. Our approach explicitly encodes observations into a volumetric representation that enables amortized rendering. We demonstrate that although continuous radiance field representations have gained a lot of attention due to their expressive power, our simple approach obtains comparable or even better novel view reconstruction quality comparing with state-of-the-art baselines while increasing rendering speed by over 400x. Our model is trained in a category-agnostic manner and does not require scene-specific optimization. Therefore, it is able to generalize novel view synthesis to object categories not seen during training. In addition, we show that with our simple formulation, we can use view synthesis as a self-supervision signal for efficient learning of 3D geometry without explicit 3D supervision.

****

Network Generalization Prediction for Safety Critical Tasks in Novel Operating Domains
Molly O'Brien, Mike Medoff, Julia Bukowski, Gregory D. Hager; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 614-622

It is well known that Neural Network (network) performance often degrades when a network is used in novel operating domains that differ from its training and testing domains. This is a major limitation, as networks are being integrated into safety critical, cyber-physical systems that must work in unconstrained environ ments, e.g., perception for autonomous vehicles. Training networks that generalize to novel operating domains and that extract robust features is an active area of research, but previous work fails to predict what the network performance will be in novel operating domains. We propose the task Network Generalization Prediction: predicting the expected network performance in novel operating domains. We describe the network performance in terms of an interpretable Context Subspace, and we propose a methodology for selecting the features of the Context Subspace that provide the most information about the network performance. We identify the Context Subspace for a pretrained Faster RCNN network performing pedestrian detection on the Berkeley Deep Drive (BDD) Dataset, and demonstrate Network Generalization Prediction accuracy within 5% of observed performance. We also demonstrate that the Context Subspace from the BDD Dataset is informative for completely unseen datasets, JAAD and Cityscapes, where predictions have a bias of 10% or less.

****

Generalized Clustering and Multi-Manifold Learning With Geometric Structure Preservation

Lirong Wu, Zicheng Liu, Jun Xia, Zelin Zang, Siyuan Li, Stan Z. Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 139-147

Though manifold-based clustering has become a popular research topic, we observe that one important factor has been omitted by these works, namely that the defined clustering loss may corrupt the local and global structure of the latent space. In this paper, we propose a novel Generalized Clustering and Multi-manifold Learning (GCML) framework with geometric structure preservation for generalized data, i.e., not limited to 2-D image data and has a wide range of applications in speech, text, and biology domains. In the proposed framework, manifold clustering is done in the latent space guided by a clustering loss. To overcome the problem that the clustering-oriented loss may deteriorate the geometric structure of the latent space, an isometric loss is proposed for preserving intra-manifold structure locally and a ranking loss for inter-manifold structure globally. Extensive experimental results have shown that GCML exhibits superior performance to counterparts in terms of qualitative visualizations and quantitative metrics, which demonstrates the effectiveness of preserving geometric structure.

****

Batch Normalization Tells You Which Filter Is Important

Junghun Oh, Heewon Kim, Sungyong Baik, Cheeun Hong, Kyoung Mu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2645-2654

The goal of filter pruning is to search for unimportant filters to remove in order to make convolutional neural networks (CNNs) efficient without sacrificing the performance in the process. The challenge lies in finding information that can help determine how important or relevant each filter is with respect to the final output of neural networks. In this work, we share our observation that the batch normalization (BN) parameters of pre-trained CNNs can be used to estimate the feature distribution of activation outputs, without processing of training data. Upon observation, we propose a simple yet effective filter pruning method by evaluating the importance of each filter based on the BN parameters of pre-trained CNNs. The experimental results on CIFAR-10 and ImageNet demonstrate that the proposed method can achieve outstanding performance with and without fine-tuning in terms of the trade-off between the accuracy drop and the reduction in computational complexity and number of parameters of pruned networks.

****

Sandwich Batch Normalization: A Drop-In Replacement for Feature Distribution Heterogeneity

Xinyu Gong, Wuyang Chen, Tianlong Chen, Zhangyang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2494-2504

We present Sandwich Batch Normalization (SaBN), a frustratingly easy improvement of Batch Normalization (BN) with only a few lines of code changes. SaBN is moti vated by addressing the inherent feature distribution heterogeneity that one can be identified in many tasks, which can arise from data heterogeneity (multiple input domains) or model heterogeneity (dynamic architectures, model conditioning, etc.). Our SaBN factorizes the BN affine layer into one shared sandwich affine layer, cascaded by several parallel independent affine layers. Concrete analysis reveals that, during optimization, SaBN promotes balanced gradient norms while still preserving diverse gradient directions -- a property that many application tasks seem to favor. We demonstrate the prevailing effectiveness of SaBN as a drop-in replacement in four tasks: conditional image generation, neural architecture search (NAS), adversarial training, and arbitrary style transfer. Leveraging SaBN immediately achieves better Inception Score and FID on CIFAR-10 and ImageNet conditional image generation with three state-of-the-art GANs; boosts the performance of a state-of-the-art weight-sharing NAS algorithm significantly on NAS-Bench-201; substantially improves the robust and standard accuracies for adversarial defense; and produces superior arbitrary stylized results. We also provide visualizations and analysis to help understand why SaBN works. Codes are available at: https://github.com/VITA-Group/Sandwich-Batch-Normalization.

****

V-SlowFast Network for Efficient Visual Sound Separation

Lingyu Zhu, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1464-1474

The objective of this paper is to perform visual sound separation: i) we study visual sound separation on spectrograms of different temporal resolutions; ii) we propose a new light yet efficient three-stream framework V-SlowFast that operates on Visual frame, Slow spectrogram, and Fast spectrogram. The Slow spectrogram captures the coarse temporal resolution while the Fast spectrogram contains the fine-grained temporal resolution; iii) we introduce two contrastive objectives to encourage the network to learn discriminative visual features for separating sounds; iv) we propose an audio-visual global attention module for audio and visual feature fusion; v) the introduced V-SlowFast model outperforms previous state-of-the-art in single-frame based visual sound separation on small- and large-scale datasets: MUSIC-21, AVE, and VGG-Sound. We also propose a small V-SlowFast architecture variant, which achieves 74.2% reduction in the number of model parameters and 81.4% reduction in GMACs compared to the previous multi-stage models. Project page: https://ly-zhu.github.io/V-SlowFast

****

Online Knowledge Distillation by Temporal-Spatial Boosting

Chengcheng Li, Zi Wang, Hairong Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 197-206

Online knowledge distillation (KD) mutually trains a group of student networks from scratch in a peer-teaching manner, eliminating the need for pre-trained teacher models. However, supervision from peers can be noisy, especially in the early stage of training. In this paper, we propose a novel method for online knowledge distillation by temporal-spatial boosting (TSB). The proposed method constructs superior "teachers" with two modules, temporal accumulator and spatial integrator. Specifically, the temporal accumulator leverages the previous outputs of networks during training and produces a representative prediction over all classes. Instead of merely imitating the outputs of other networks as in vanilla online KD, we further propose the so-called spatial integrator that consolidates the knowledge learned by all networks and yields a stronger instructor. The operations of these two modules are simple and straightforward, which can be computed efficiently on the fly during training. The proposed method can improve the efficiency of transferring effective knowledge as well as stabilize the training process. Experimental results on various benchmark datasets and network structures validate the effectiveness of the proposed method over the state-of-the-art.

****

Learning To Generate the Unknowns as a Remedy to the Open-Set Domain Shift

Mahsa Baktashmotlagh, Tianle Chen, Mathieu Salzmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 207-216

In many situations, the data one has access to at test time follows a different distribution from the training data. Over the years, this problem has been tackled by closed-set domain adaptation techniques. Recently, open-set domain adaptation has emerged to address the more realistic scenario where additional unknown classes are present in the target data. In this setting, existing techniques focus on the challenging task of isolating the unknown target samples, so as to avoid the negative transfer resulting from aligning the source feature distributions with the broader target one that encompasses the additional unknown classes. Here, we propose a simpler and more effective solution consisting of complementing the source data distribution and making it comparable to the target one by enabling the model to generate source samples corresponding to the unknown target classes. We formulate this as a general module that can be incorporated into any existing closed-set approach and show that this strategy allows us to outperform the state-of-the-art on open-set domain adaptation benchmark datasets.

****

Mutual Learning of Joint and Separate Domain Alignments for Multi-Source Domain Adaptation

Yuanyuan Xu, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1890-1899

Multi-Source Domain Adaptation (MSDA) aims at transferring knowledge from multiple labeled source domains to benefit the task in an unlabeled target domain. The challenges of MSDA lie in mitigating domain gaps and combining information from diverse source domains. In most existing methods, the multiple source domains can be jointly or separately aligned to the target domain. In this work, we consider that these two types of methods, i.e. joint and separate domain alignments, are complementary and propose a mutual learning based alignment network (MLAN) to combine their advantages. Specifically, our proposed method is composed of three components, i.e. a joint alignment branch, a separate alignment branch, and a mutual learning objective between them. In the joint alignment branch, the samples from all source domains and the target domain are aligned together, with a single domain alignment goal, while in the separate alignment branch, each source domain is individually aligned to the target domain. Finally, by taking advantage of the complementarity of joint and separate domain alignment mechanisms, mutual learning is used to make the two branches learn collaboratively. Compared with other existing methods, our proposed MLAN integrates information of different domain alignment mechanisms and thus can mine rich knowledge from multiple domains for better performance. The experiments on DomainNet, Office-31, and Digits-five datasets demonstrate the effectiveness of our method.

****

Deep Online Fused Video Stabilization

Zhenmei Shi, Fuhao Shi, Wei-Sheng Lai, Chia-Kai Liang, Yingyu Liang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1250-1258

We present a deep neural network (DNN) that uses both sensor data (gyroscope) and image content (optical flow) to stabilize videos through unsupervised learning. The network fuses optical flow with real/virtual camera pose histories into a joint motion representation. Next, the LSTM cell infers the new virtual camera pose, which is used to generate a warping grid that stabilizes the video frames. We adopt a relative motion representation as well as a multi-stage training strategy to optimize our model without any supervision. To the best of our knowledge, this is the first DNN solution that adopts both sensor data and image content for video stabilization. We validate the proposed framework through ablation studies and demonstrate that the proposed method outperforms the state-of-art alternative solutions via quantitative evaluations and a user study. Check out our video results, code and dataset at our website.

****

Parsing Line Chart Images Using Linear Programming

Hajime Kato, Mitsuru Nakazawa, Hsuan-Kung Yang, Mark Chen, Björn Stenger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2109-2118

This paper proposes a method for automatically recovering data from chart images

. In particular we focus on the task of estimating line charts, as the most common chart type, in a fully automatic way that handles line occlusions, as well as lines of different styles, e.g., dashed or dotted. For this, we first train a single semantic segmentation network to predict probability maps for each different line styles. We then construct a graph based on this output and formulate the line tracing task as a minimum-cost-flow problem, optimizing a cost function using linear programming. From the traced lines, the axes, and text labels, we recover the numerical values used to generate the chart. In experiments on six datasets, containing both synthesized and crawled images, we show significant improvements over prior work.
****
TricubeNet: 2D Kernel-Based Object Representation for Weakly-Occluded Oriented Object Detection
Beomyoung Kim, Janghyeon Lee, Sihaeng Lee, Doyeon Kim, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 167-176
We present a novel approach for oriented object detection, named TricubeNet, which localizes oriented objects using visual cues (i.e., heatmap) instead of oriented box offsets regression. We represent each object as a 2D Tricube kernel and extract bounding boxes using simple image-processing algorithms. Our approach is able to (1) obtain well-arranged boxes from visual cues, (2) solve the angle discontinuity problem, and (3) can save computational complexity due to our anchor-free modeling. To further boost the performance, we propose some effective techniques for size-invariant loss, reducing false detections, extracting rotation-invariant features, and heatmap refinement. To demonstrate the effectiveness of our TricubeNet, we experiment on various tasks for weakly-occluded oriented object detection: detection in an aerial image, densely packed object image, and text image. The extensive experimental results show that our TricubeNet is quite effective for oriented object detection. Code is available at https://github.com/qjadud1994/TricubeNet.
****
CrossLocate: Cross-Modal Large-Scale Visual Geo-Localization in Natural Environments Using Rendered Modalities
Jan Tomešek, Martin ■adík, Jan Brejcha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3174-3183
We propose a novel approach to visual geo-localization in natural environments. This is a challenging problem due to vast localization areas, the variable appearance of outdoor environments and the scarcity of available data. In order to make the research of new approaches possible, we first create two databases containing "synthetic" images of various modalities. These image modalities are rendered from a 3D terrain model and include semantic segmentations, silhouette maps and depth maps. By combining the rendered database views with existing datasets of photographs (used as "queries" to be localized), we create a unique benchmark for visual geo-localization in natural environments, which contains correspondences between query photographs and rendered database imagery. The distinct ability to match photographs to synthetically rendered databases defines our task as "cross-modal". On top of this benchmark, we provide thorough ablation studies analysing the localization potential of the database image modalities. We reveal the depth information as the best choice for outdoor localization. Finally, based on our observations, we carefully develop a fully-automatic method for large-scale cross-modal localization using image retrieval. We demonstrate its localization performance outdoors in the entire state of Switzerland. Our method reveals a large gap between operating within a single image domain (e.g. photographs) and working across domains (e.g. photographs matched to rendered images), as gained knowledge is not transferable between the two. Moreover, we show that modern localization methods fail when applied to such a cross-modal task and that our method achieves significantly better results than state-of-the-art approaches. The datasets, code and trained models are available on the project website: http://cphoto.fit.vutbr.cz/crosslocate/.
****

Symmetric-Light Photometric Stereo

Kazuma Minami, Hiroaki Santo, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2706-2714

This paper presents symmetric-light photometric stereo for surface normal estimation, in which directional lights are distributed symmetrically with respect to the optic center. Unlike previous studies of ring-light settings that required the information of ring radius, we show that even without the knowledge of the exact light source locations or their distances from the optic center, the symmetric configuration provides us sufficient information for recovering unique surface normals without ambiguity. Specifically, under the symmetric lights, measurements of a pair of scene points having distinct surface normals but the same albedo yield a system of constrained quadratic equations about the surface normal, which has a unique solution. Experiments demonstrate that the proposed method alleviates the need for geometric light source calibration while maintaining the accuracy of calibrated photometric stereo.

****

CharacterGAN: Few-Shot Keypoint Character Animation and Reposing

Tobias Hinz, Matthew Fisher, Oliver Wang, Eli Shechtman, Stefan Wermter; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1988-1997

We introduce CharacterGAN, a generative model that can be trained on only a few samples (8 - 15) of a given character. Our model generates novel poses based on keypoint locations, which can be modified in real time while providing interactive feedback, allowing for intuitive reposing and animation. Since we only have very limited training samples, one of the key challenges lies in how to address (dis)occlusions, e.g. when a hand moves behind or in front of a body. To address this, we introduce a novel layering approach which explicitly splits the input keypoints into different layers which are processed independently. These layers represent different parts of the character and provide a strong implicit bias that helps to obtain realistic results even with strong (dis)occlusions. To combine the features of individual layers we use an adaptive scaling approach conditioned on all keypoints. Finally, we introduce a mask connectivity constraint to reduce distortion artifacts that occur with extreme out-of-distribution poses at test time. We show that our approach outperforms recent baselines and creates realistic animations for diverse characters. We also show that our model can handle discrete state changes, for example a profile facing left or right, that the different layers do indeed learn features specific for the respective keypoints in those layers, and that our model scales to larger datasets when more data is available. Code is available at https://github.com/tohinz/CharacterGAN.

****

Boosting Contrastive Self-Supervised Learning With False Negative Cancellation

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, Maryam Khademi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2785-2795

Self-supervised representation learning has made significant leaps fueled by progress in contrastive learning, which seeks to learn transformations that embed positive input pairs nearby, while pushing negative pairs far apart. While positive pairs can be generated reliably (e.g., as different views of the same image), it is difficult to accurately establish negative pairs, defined as samples from different images regardless of their semantic content or visual features. A fundamental problem in contrastive learning is mitigating the effects of false negatives. Contrasting false negatives induces two critical issues in representation learning: discarding semantic information and slow convergence. In this paper, we propose novel approaches to identify false negatives, as well as two strategies to mitigate their effect, i.e. false negative elimination and attraction, while systematically performing rigorous evaluations to study this problem in detail. Our method exhibits consistent improvements over existing contrastive learning-based methods. Without labels, we identify false negatives with 40% accuracy among 1000 semantic classes on ImageNet, and achieve 5.8% absolute improvement i n top-1 accuracy over the previous state-of-the-art when finetuning with 1% labels.

****

FASSST: Fast Attention Based Single-Stage Segmentation Net for Real-Time Instance Segmentation

Yuan Cheng, Rui Lin, Peining Zhen, Tianshu Hou, Chiu Wa Ng, Hai-Bao Chen, Hao Yu, Ngai Wong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2210-2218

Real-time instance segmentation is crucial in various AI applications. This work designs a network named Fast Attention based Single-Stage Segmentation NeT (FASSST) that performs instance segmentation with video-grade speed. Using an instance attention module (IAM), FASSST quickly locates target instances and segments with region of interest (ROI) feature fusion (RFF) aggregating ROI features from pyramid mask layers. The module employs an efficient single-stage feature regression, straight from features to instance coordinates and class probabilities. Experiments on COCO and CityScapes datasets show that FASSST achieves state-of-the-art performance under competitive accuracy: real-time inference of 47.5FPS on a GTX1080Ti GPU and 5.3FPS on a Jetson Xavier NX board with only 71.6GFLOPs.

****

InfographicVQA

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1697-1706

Infographics communicate information using a combination of textual, graphical and visual elements. This work explores the automatic understanding of infographic images by using a Visual Question Answering technique. To this end, we present InfographicVQA, a new dataset comprising a diverse collection of infographics and question-answer annotations. The questions require methods that jointly reason over the document layout, textual content, graphical elements, and data visualizations. We curate the dataset with an emphasis on questions that require elementary reasoning and basic arithmetic skills. For VQA on the dataset, we evaluate two Transformer-based strong baselines. Both the baselines yield unsatisfactory results compared to near perfect human performance on the dataset. The results suggest that VQA on infographics--images that are designed to communicate information quickly and clearly to human brain--is ideal for benchmarking machine understanding of complex document images. The dataset is available for download at docvqa.org

****

No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency

S. Alireza Golestaneh, Saba Dadsetan, Kris M. Kitani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1220-1230

The goal of No-Reference Image Quality Assessment (NR-IQA) is to estimate the perceptual image quality in accordance with subjective evaluations, it is a complex and unsolved problem due to the absence of the pristine reference image. In this paper, we propose a novel model to address the NR-IQA task by leveraging a hybrid approach that benefits from Convolutional Neural Networks (CNNs) and self-attention mechanism in Transformers to extract both local and non-local features from the input image. We capture local structure information of the image via CNNs, then to circumvent the locality bias among the extracted CNNs features and obtain a non-local representation of the image, we utilize Transformers on the extracted features where we model them as a sequential input to the Transformer model. Furthermore, to improve the monotonicity correlation between the subjective and objective scores, we utilize the relative distance information among the images within each batch and enforce the relative ranking among them. Last but not least, we observe that the performance of NR-IQA models degrades when we apply equivariant transformations (e.g. horizontal flipping) to the inputs. Therefore, we propose a method that leverages self-consistency as a source of self-supervision to improve the robustness of NR-IQA models. Specifically, we enforce self-c onsistency between the outputs of our quality assessment model for each image and its transformation (horizontally flipped) to utilize the rich self-supervisory information and reduce the uncertainty of the model. To demonstrate the effectiveness of our work, we evaluate it on seven standard IQA datasets (both synthetic and authentic) and show that our model achieves state-of-the-art results on various datasets.

****

Attribute-Based Deep Periocular Recognition: Leveraging Soft Biometrics to Improve Periocular Recognition

Veeru Talreja, Nasser M. Nasrabadi, Matthew C. Valenti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4041-4050

In recent years, periocular recognition has been developed as a valuable biometric identification approach, especially in wild environments (for example, masked faces due to COVID-19 pandemic) where facial recognition may not be applicable. This paper presents a new deep periocular recognition framework called attribute-based deep periocular recognition (ADPR), which predicts soft biometrics and incorporates the prediction into a periocular recognition algorithm to determine identity from periocular images with high accuracy. We propose an end-to-end framework, which uses several shared convolutional neural network (CNN) layers (a common network) whose output feeds two separate dedicated branches (modality dedicated layers); the first branch classifies periocular images while the second branch predicts soft biometrics. Next, the features from these two branches are fused together for a final periocular recognition. The proposed method is different from existing methods as it not only uses a shared CNN feature space to train these two tasks jointly, but it also fuses predicted soft biometric features with the periocular features in the training step to improve the overall periocular recognition performance. Our proposed model is extensively evaluated using four different publicly available datasets. Experimental results indicate that our soft biometric based periocular recognition approach outperforms other state-of-the-art methods for periocular recognition in wild environments.

****

Coupled Training for Multi-Source Domain Adaptation

Ohad Amosy, Gal Chechik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 420-429

Unsupervised domain adaptation is often addressed by learning a joint representation of labeled samples from a source domain and unlabeled samples from a target domain. Unfortunately, hard sharing of representation may hurt adaptation because of negative transfer, where features that are useful for source domains are learned even if they hurt inference on the target domain. Here, we propose an alternative, soft sharing scheme. We train separate but weakly-coupled models for the source and the target data, while encouraging their predictions to agree. Training the two coupled models jointly effectively exploits the distribution over unlabeled target data and achieves high accuracy on the target. Specifically, we show analytically and empirically that the decision boundaries of the target model converge to low-density "valleys" of the target distribution. We evaluate our approach on four multi-source domain adaptation (MSDA) benchmarks, digits, amazon text reviews, Office-Caltech, and images (DomainNet). We find that it consistently outperforms current MSDA SoTA, sometimes by a very large margin.

****

Compressed Sensing MRI Reconstruction With Co-VeGAN: Complex-Valued Generative Adversarial Network

Bhavya Vasudeva, Puneesh Deora, Saumik Bhattacharya, Pyari Mohan Pradhan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 672-681

Compressed sensing (CS) is extensively used to reduce magnetic resonance imaging (MRI) acquisition time. State-of-the-art deep learning-based methods have proven effective in obtaining fast, high-quality reconstruction of CS-MR images. However, they treat the inherently complex-valued MRI data as real-valued entities by extracting the magnitude content or concatenating the complex-valued data as t wo real-valued channels for processing. In both cases, the phase content is discarded. To address the fundamental problem of real-valued deep networks, i.e. their inability to process complex-valued data, we propose a complex-valued generative adversarial network (Co-VeGAN) framework, which is the first-of-its-kind generative model exploring the use of complex-valued weights and operations. Further, since real-valued activation functions do not generalize well to the complex-valued space, we propose a novel complex-valued activation function that is sensitive to the input phase and has a learnable profile. Extensive evaluation of the proposed approach on different datasets demonstrates that it significantly outperforms the existing CS-MRI reconstruction techniques.

****

Uncertainty Learning Towards Unsupervised Deformable Medical Image Registration

Xuan Gong, Luckyson Khaidem, Wentao Zhu, Baochang Zhang, David Doermann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2484-2493

Uncertainty estimation in medical image registration enables surgeons to evaluate the operative risk based on the trustworthiness of the registered image data thus of paramount importance for practical clinical applications. Despite the recent promising results obtained with deep unsupervised learning-based registration methods, reasoning about uncertainty of unsupervised registration models remains largely unexplored. In this work, we propose a predictive module to learn the registration and uncertainty in correspondence simultaneously. Our framework introduces empirical randomness and registration error based uncertainty prediction. We systematically assess the performances on two MRI datasets with different ensemble paradigms. Experimental results highlight that our proposed framework significantly improves the registration accuracy and uncertainty compared with the baseline.

****

Video and Text Matching With Conditioned Embeddings

Ameen Ali, Idan Schwartz, Tamir Hazan, Lior Wolf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1565-1574

We present a method for matching a text sentence from a given corpus to a given video clip and vice versa. Traditionally video and text matching is done by learning a shared embedding space and the encoding of one modality is independent of the other. In this work, we encode the dataset data in a way that takes into account the query's relevant information. The power of the method is demonstrated to arise from pooling the interaction data between words and frames. Since the encoding of the video clip depends on the sentence compared to it, the representation needs to be recomputed for each potential match. To this end, we propose an efficient shallow neural network. Its training employs a hierarchical triplet loss that is extendable to paragraph/video matching. The method is simple, provide explainability, and achieves a state-of-the-art-results, for both sentence-clip and video-text by a sizable margin across five different datasets: ActivityNet, DiDeMo, YouCook2, MSR-VTT, and LSMDC. We also show that our conditioned representation can be transferred to video-guided machine translation, where we improved the current results on VATEX. Source code is available at https://github.com/AmeenAli/VideoMatch.

****

Multi-Head Deep Metric Learning Using Global and Local Representations

Mohammad K. Ebrahimpour, Gang Qian, Allison Beach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3031-3040

Deep Metric Learning (DML) aims to learn a data embedding space in which similar data points are grouped together while dissimilar data points are pushed away from each other. Successful DML models often require strong local and global representations, however, effective integration of local and global features in DML model training is a challenge. DML models are often trained with specific loss functions, including pairwise-based and proxy-based losses. The pairwise-based loss functions leverage rich semantic relations among data points, however, they often suffer from slow convergence during DML model training. On the other hand, the proxy-based loss functions often lead to significant speedups in convergence during training, while the rich relations among data points are often not fully explored by the proxy-based losses. In this paper, we propose a novel DML approach to address these challenges. The proposed DML approach makes use of a hybrid loss by integrating the pairwise-based and the proxy-based loss functions to leverage rich data-to-data relations as well as fast convergence. Furthermore, the proposed DML approach utilizes both global and local features to obtain rich representations in DML model training. Finally, We also use the second-order attention for feature enhancement to improve accurate and efficient retrieval. In our experiments, we extensively evaluated the proposed DML approach on four public benchmarks, and the experimental results demonstrate that the proposed method achieved state-of-the-art performance on all benchmarks, often with a large margin.

****

REGroup: Rank-Aggregating Ensemble of Generative Classifiers for Robust Predictions

Lokender Tiwari, Anish Madan, Saket Anand, Subhashis Banerjee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2595-2604

Deep Neural Networks (DNNs) are often criticized for being susceptible to adversarial attacks. Most successful defense strategies adopt adversarial training or random input transformations that typically require retraining or fine-tuning the model to achieve reasonable performance. In this work, our investigations of intermediate representations of a pre-trained DNN lead to an interesting discovery pointing to intrinsic robustness to adversarial attacks. We find that we can learn a generative classifier by statistically characterizing the neural response of an intermediate layer to clean training samples. The predictions of multiple such intermediate-layer based classifiers, when aggregated, show unexpected robustness to adversarial attacks. Specifically, we devise an ensemble of these generative classifiers that rank-aggregates their predictions via a Borda count-based consensus. Our proposed approach uses a subset of the clean training data and a pre-trained model, and yet is agnostic to network architectures or the adversarial attack generation method. We show extensive experiments to establish that our defense strategy achieves state-of-the-art performance on the ImageNet validation set.

****

Learning Foreground-Background Segmentation From Improved Layered GANs

Yu Yang, Hakan Bilen, Qiran Zou, Wing Yin Cheung, Xiangyang Ji; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2524-2533

Deep learning approaches heavily rely on high-quality human supervision which is nonetheless expensive, time-consuming, and error-prone, especially for image segmentation task. In this paper, we propose a method to automatically synthesize paired photo-realistic images and segmentation masks for the use of training a foreground-background segmentation network. In particular, we learn a generative adversarial network that decomposes an image into foreground and background layers, and avoid trivial decompositions by maximizing mutual information between generated images and latent variables. The improved layered GANs can synthesize higher quality datasets from which segmentation networks of higher performance can be learned. Moreover, the segmentation networks are employed to stabilize the training of layered GANs in return, which are further alternately trained with Layered GANs. Experiments on a variety of single-object datasets show that our method achieves competitive generation quality and segmentation performance compared to related methods.

****

HHP-Net: A Light Heteroscedastic Neural Network for Head Pose Estimation With Uncertainty

Giorgio Cantarini, Federico Figari Tomenotti, Nicoletta Noceti, Francesca Odone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3521-3530

In this paper we introduce a novel method to estimate the head pose of people in

single images starting from a small set of head keypoints. To this purpose, we propose a regression model that exploits keypoints computed automatically by 2D pose estimation algorithms and outputs the head pose represented by yaw, pitch, and roll. Our model is simple to implement and more efficient with respect to the state of the art -- faster in inference and smaller in terms of memory occupancy -- with comparable accuracy. Our method also provides a measure of the heteroscedastic uncertainties associated with the three angles, through an appropriately designed loss function; we show there is a correlation between error and uncertainty values, thus this extra source of information may be used in subsequent computational steps. As an example application, we address social interaction analysis in images: we propose an algorithm for a quantitative estimation of the level of interaction between people, starting from their head poses and reasoning on their mutual positions.
****
Less Can Be More: Sound Source Localization With a Classification Model
Arda Senocak, Hyeonggon Ryu, Junsik Kim, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3308-3317
In this paper, we tackle sound localization as a natural outcome of the audio-visual video classification problem. Differently from the existing sound localization approaches, we do not use any explicit sub-modules or training mechanisms but use simple cross-modal attention on top of the representations learned by a classification loss. Our key contribution is to show that a simple audio-visual classification model has the ability to localize sound sources accurately and to give on par performance with state-of-the-art methods by proving that indeed "less is more". Furthermore, we propose potential applications that can be built based on our model. First, we introduce informative moment selection to enhance the localization task learning in the existing approaches compare to mid-frame usage. Then, we introduce a pseudo bounding box generation procedure that can significantly boost the performance of the existing methods in semi-supervised settings or be used for large-scale automatic annotation with minimal effort from any video dataset.
****
One-Class Learned Encoder-Decoder Network With Adversarial Context Masking for Novelty Detection
John Taylor Jewell, Vahid Reza Khazaie, Yalda Mohsenzadeh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3591-3601
Novelty detection is the task of recognizing samples that do not belong to the distribution of the target class. During training, the novelty class is absent, preventing the use of traditional classification approaches. Deep autoencoders have been widely used as a base of many novelty detection methods. In particular, context autoencoders have been successful in the novelty detection task because of the more effective representations they learn by reconstructing original images from randomly masked images. However, a significant drawback of context autoencoders is that random masking fails to consistently cover important structures of the input image, leading to suboptimal representations - especially for the novelty detection task. In this paper, to optimize input masking, we introduce a Mask Module that learns to generate optimal masks and a Reconstructor that aims to reconstruct masked images. The networks are trained in an adversarial setting in which the Mask Module seeks to maximize the reconstruction error that the Reconstructor is minimizing. When applied to novelty detection, the proposed approach learns semantically richer representations compared to context autoencoders and enhances novelty detection at test time through more optimal masking. Novelty detection experiments on the MNIST and CIFAR-10 image datasets demonstrate the proposed approach's superiority over cutting-edge methods. In a further experiment on the UCSD video dataset for novelty detection, the proposed approach achieves a frame-level Area Under the Curve (AUC) of 99.02% and an Equal Error Rate (EER) of 5.4%, exceeding recent state-of-the-art models.
****

Face Verification With Challenging Imposters and Diversified Demographics

Adrian Popescu, Liviu-Daniel ■tefan, Jérôme Deshayes-Chossart, Bogdan Ionescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3357-3366

Face verification aims to distinguish between genuine and imposter pairs of faces, which include the same or different identities, respectively. The performance reported in recent years gives the impression that the task is practically solved. Here, we revisit the problem and argue that existing evaluation datasets were built using two oversimplifying design choices. First, the usual identity selection to form imposter pairs is not challenging enough because, in practice, verification is needed to detect challenging imposters. Second, the underlying demographics of existing datasets are often insufficient to account for the wide diversity of facial characteristics of people from across the world. To mitigate these limitations, we introduce the FaVCI2D dataset. Imposter pairs are challenging because they include visually similar faces selected from a large pool of demographically diversified identities. The dataset also includes metadata related to gender, country and age to facilitate fine-grained analysis of results. FaVCI2D is generated from freely distributable resources and is compliant with data protection regulations. Experiments with state-of-the-art deep models that provide nearly 100% performance on existing datasets show a significant performance drop for FaVCI2D, confirming our starting hypothesis. Equally important, we analyze legal and ethical challenges which appeared in recent years and hindered the development of face analysis research. We introduce a series of design choices which address these challenges and make the dataset constitution and usage more sustainable and fairer.

****

Leveraging Test-Time Consensus Prediction for Robustness Against Unseen Noise

Anindya Sarkar, Anirban Sarkar, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1839-1848

We propose a method to improve DNN robustness against unseen noisy corruptions, such as Gaussian noise, Shot Noise, Impulse Noise, Speckle noise with different levels of severity by leveraging ensemble technique through a consensus based prediction method using self-supervised learning at inference time. We also propose to enhance the model training by considering other aspects of the issue i.e. noise in data and better representation learning which shows even better generalization performance with the consensus based prediction strategy. We report results of each noisy corruption on the standard CIFAR10-C and ImageNet-C benchmark which shows significant boost in performance over previous methods. We also introduce results for MNIST-C and TinyImagenet-C to show usefulness of our method across datasets of different complexities to provide robustness against unseen noise. We show results with different architectures to validate our method against other baseline methods, and also conduct experiments to show the usefulness of each part of our method.

****

Detecting Tear Gas Canisters With Limited Training Data

Ashwin D'Cruz, Christopher Tegho, Sean Greaves, Lachlan Kermode; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3674-3682

Human rights investigations often require triaging large volumes of open source data in order to find moments within image, or video that are relevant to a given investigation and warrant further inspection. Searching for images of tear gas usage online manually is laborious and time-consuming. In this paper, we focus on object detection models to facilitate discovery and identification of tear gas canisters for human rights monitors. For CNN based object detection to work, a large amount of training data is required, and prior to our work, a dataset of tear gas canisters did not exist. To achieve our objective, we benchmark methods for training object detectors using limited labelled data: we fine-tune different object detection models on the limited labelled data and compare performance to a few shot detector and augmentation strategies using synthetic data. We prov ide a dataset for evaluating and training tear gas canister detectors and show how such detectors can be deployed for a real world application such as investigating human rights violations. Our experiments show that fine-tuning state of the art detectors perform as well as the few shot detector, and including synthetic data can improve results.

****

Improve Image Captioning by Estimating the Gazing Patterns From the Caption

Rehab Alahmadi, James Hahn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1025-1034

Recently, there has been much interest in developing image captioning models. State-of-the-art models reached a good performance in producing human-like descriptions from image features that are extracted from neural network models such as CNN and R-CNN. However, none of the previous methods have encapsulated explicit features that reflect a human perception of the images such as gazing patterns without the use of the eye-tracking systems. In this paper, we hypothesize that the nouns (i.e. entities) and their orders in the image description reflect human gazing patterns and perception. To this end, we estimate the sequence of the gazed objects from the words in the captions and then train a pointer network to learn to produce such sequence automatically given a set of objects in new images. We incorporate the suggested sequence by pointer network in existing image caption models and investigate its performance. Our experiments show a significant increase in the performance of the image captioning models when the sequence of the gazed objects are utilized as additional features (up to 13 points improvement in CIDEr score when combined with Neural Image Caption model).

****

Consistent Cell Tracking in Multi-Frames With Spatio-Temporal Context by Object-Level Warping Loss

Junya Hayashida, Kazuya Nishimura, Ryoma Bise; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1727-1736

Multi-object tracking is essential in biomedical image analysis. Most multi-object tracking methods follow a tracking-by-detection approach that involves using object detectors and learning the appearance feature models of the detected regions for association. Although these methods can learn the appearance similarity features to identify the same objects among frames, they have difficulties identifying the same cells because cells have a similar appearance and their shapes change as they migrate. In addition, cells often partially overlap for several frames. In this case, even an expert biologist would require knowledge of the spatial-temporal context in order to identify individual cells. To tackle such difficult situations, we propose a cell-tracking method that can effectively use the spatial-temporal context in multiple frames by using long-term motion estimation and an object-level warping loss. We conducted experiments showing that the proposed method outperformed state-of-the-art methods under various conditions on real biological images.

****

Learnable Adaptive Cosine Estimator (LACE) for Image Classification

Joshua Peeples, Connor H. McCurley, Sarah Walker, Dylan Stewart, Alina Zare; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3479-3489

In this work, we propose a new loss to improve feature discriminability and classification performance. Motivated by the adaptive cosine/coherence estimator (ACE), our proposed method incorporates angular information that is inherently learned by artificial neural networks. Our learnable ACE (LACE) transforms the data into a new "whitened" space that improves the inter-class separability and intra-class compactness. We compare our LACE to alternative state-of-the art softmax-based and feature regularization approaches. Our results show that the proposed method can serve as a viable alternative to cross entropy and angular softmax approaches. Our code is publicly available.

****

Auto-X3D: Ultra-Efficient Video Understanding via Finer-Grained Neural Architecture Search

Yifan Jiang, Xinyu Gong, Junru Wu, Humphrey Shi, Zhicheng Yan, Zhangyang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2554-2563

Efficient video architecture is the key to the deployment of video action recognition systems on devices with limited computing capabilities. Unfortunately, existing video architectures are often computationally intensive and not suitable for such applications. The recent X3D work presents a new family of efficient video models by expanding a hand-crafted image architecture along multiple axes, such as space, time, width, and depth. Although operating in a conceptually large space, X3D searched one axis at a time, and merely explored a small set of 30 architectures in total, which does not sufficiently explore the space. This paper bypasses existing 2D architectures, and directly searched for 3D architectures in a fine-grained space, where block type, filter number, expansion ratio and attention block are jointly searched. A probabilistic neural architecture search method is adopted to efficiently search in such a large space. Evaluations on Kinetics and Something-Something-V2 benchmarks confirm our \autoxthreed models outperform existing ones in accuracy up to 1.7% under similar FLOPs, and reduce the computational cost up to 1.74 times to reach similar performance. Code will be publicly available.

****

Learned Event-Based Visual Perception for Improved Space Object Detection

Nikolaus Salvatore, Justin Fletcher; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2888-2897

The detection of dim artificial Earth satellites using ground-based electro-optical sensors, particularly in the presence of background light, is technologically challenging. This perceptual task is foundational to our understanding of the space environment, and grows in importance as the number, variety, and dynamism of space objects increases. We present a hybrid image- and event-based architecture that leverages dynamic vision sensing technology to detect resident space objects in geosynchronous Earth orbit. Given the asynchronous, one-dimensional image data supplied by a dynamic vision sensor, our architecture applies conventional image feature extractors to integrated, two-dimensional frames in conjunction with point-cloud feature extractors, such as PointNet, in order to increase detection performance for dim objects in scenes with high background activity. In addition, an end-to-end event-based imaging simulator is developed to both produce data for model training as well as approximate the optimal sensor parameters for event-based sensing in the context of electro-optical telescope imagery. Experimental results confirm that the inclusion of point-cloud feature extractors increases recall for dim objects in the high-background regime.

****

Biomass Prediction With 3D Point Clouds From LiDAR

Liyuan Pan, Liu Liu, Anthony G. Condon, Gonzalo M. Estavillo, Robert A. Coe, Geoff Bull, Eric A. Stone, Lars Petersson, Vivien Rolland; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1330-1340

With population growth and a shrinking rural workforce, agricultural technologies have become increasingly important. Above-ground biomass (AGB) is a key trait relevant to breeding, agronomy and crop physiology field experiments. However, measuring the biomass of a cereal plot requires cutting, drying and weighing processes, which are laborious, expensive and destructive tasks. This paper proposes a non-destructive and high-throughput method to predict biomass from field samples based on Light Detection and Ranging (LiDAR). Unlike previous methods that are based on the density of a point cloud or plant height, our biomass prediction network (BioNet) additionally considers plant structure. Our BioNet contains three modules: 1) a completion module to predict missing points due to canopy occlusion; 2) a regularization module to regularize the neural representation of the whole plot; and 3) a projection module to learn the salient structures from a bird's eye view of the point cloud. An attention-based fusion block is used to achieve final biomass predictions. In addition, the complete dataset, including hand-measured biomass and LiDAR data, is made available to the community. Experime nts show that our BioNet achieves approximately 33% improvement over current state-of-the-art methods.

****

Shallow Features Guide Unsupervised Domain Adaptation for Semantic Segmentation at Class Boundaries

Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, Luigi Di Stefano; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1160-1170

Although deep neural networks have achieved remarkable results for the task of semantic segmentation, they usually fail to generalize towards new domains, especially when performing synthetic-to-real adaptation. Such domain shift is particularly noticeable along class boundaries, invalidating one of the main goals of semantic segmentation that consists in obtaining sharp segmentation masks. In this work, we specifically address this core problem in the context of Unsupervised Domain Adaptation and present a novel low-level adaptation strategy that allows us to obtain sharp predictions. Moreover, inspired by recent self-training techniques, we introduce an effective data augmentation that alleviates the noise typically present at semantic boundaries when employing pseudo-labels for self-training. Our contributions can be easily integrated into other popular adaptation frameworks, and extensive experiments show that they effectively improve performance along class boundaries.

****

S2-MLP: Spatial-Shift MLP Architecture for Vision

Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, Ping Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 297-306

Recently, visual Transformer (ViT) and its following works abandon the convolution and exploit the self-attention operation, attaining a comparable or even higher accuracy than CNN. More recently, MLP-mixer abandons both the convolution and the self-attention operation, proposing an architecture containing only MLP layers. To achieve cross-patch communications, it devises an additional token-mixing MLP besides the channel-mixing MLP. It achieves promising results when training on an extremely large-scale dataset such as JFT-300M. But it cannot achieve as outstanding performance as its CNN and ViT counterparts when training on medium-scale datasets such as ImageNet-1K and ImageNet-21K. The performance drop of MLP-mixer motivates us to rethink the token-mixing MLP. We discover that token-mixing operation in MLP-mixer is a variant of depthwise convolution with a global reception field and spatial-specific configuration. It is the global reception field and the spatial-specific property that make token-mixing MLP prone to over-fitting. In this paper, we propose a novel pure MLP architecture, spatial-shift MLP (S^2-MLP). Different from MLP-mixer, our S^2-MLP only contains channel-mixing MLP. We devise a spatial-shift operation for achieving the communication between patches. It has a local reception field and is spatial-agnostic. Meanwhile, it is parameter-free and efficient for computation. The proposed S^2-MLP attains higher recognition accuracy than MLP-mixer when training on ImageNet-1K dataset. Meanwhile, S^2-MLP accomplishes as excellent performance as ViT on ImageNet-1K dataset with considerably simpler architecture and fewer FLOPs and parameters.

****

Deep Feature Prior Guided Face Deblurring

Soo Hyun Jung, Tae Bok Lee, Yong Seok Heo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3531-3540

Most recent face deblurring methods have focused on utilizing facial shape priors such as face landmarks and parsing maps. While these priors can provide facial geometric cues effectively, they are insufficient to contain local texture details that act as important clues to solve face deblurring problem. To deal with this, we focus on estimating the deep features of pre-trained face recognition networks (e.g., VGGFace network) that include rich information about sharp faces as a prior, and adopt a generative adversarial network (GAN) to learn it. To this end, we propose a deep feature prior guided network (DFPGnet) that restores facial details using the estimated the deep feature prior from a blurred image. In our DFPGnet, the generator is divided into two streams including prior estimatio n and deblurring streams. Since the estimated deep features of the prior estimation stream are learned from the VGGFace network which is trained for face recognition not for deblurring, we need to alleviate the discrepancy of feature distributions between the two streams. Therefore, we present feature transform modules at the connecting points of the two streams. In addition, we propose a channel-attention feature discriminator and prior loss, which encourages the generator to focus on more important channels for deblurring among the deep feature prior during training. Experimental results show that our method achieves state-of-the-art performance both qualitatively and quantitatively.

****

Supervised Compression for Resource-Constrained Edge Computing Systems

Yoshitomo Matsubara, Ruihan Yang, Marco Levorato, Stephan Mandt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2685-2695

There has been much interest in deploying deep learning algorithms on low-powered devices, including smartphones, drones, and medical sensors. However, full-scale deep neural networks are often too resource-intensive in terms of energy and storage. As a result, the bulk part of the machine learning operation is therefore often carried out on an edge server, where the data is compressed and transmitted. However, compressing data (such as images) leads to transmitting information irrelevant to the supervised task. Another popular approach is to split the deep network between the device and the server while compressing intermediate features. To date, however, such split computing strategies have barely outperformed the aforementioned naive data compression baselines due to their inefficient approaches to feature compression. This paper adopts ideas from knowledge distillation and neural image compression to compress intermediate feature representations more efficiently. Our supervised compression approach uses a teacher model and a student model with a stochastic bottleneck and learnable prior for entropy coding (Entropic Student). We compare our approach to various neural image and feature compression baselines in three vision tasks and found that it achieves better supervised rate-distortion performance while maintaining smaller end-to-end latency. We furthermore show that the learned feature representations can be tuned to serve multiple downstream tasks.

****

Pose-Guided Generative Adversarial Net for Novel View Action Synthesis

Xianhang Li, Junhao Zhang, Kunchang Li, Shruti Vyas, Yogesh S. Rawat; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3861-3870

We focus on the problem of novel-view human action synthesis. Given an action video, the goal is to generate the same action from an unseen viewpoint. Naturally, novel view video synthesis is more challenging than image synthesis. It requires the synthesis of a sequence of realistic frames with temporal coherency. Besides, transferring the different actions to a novel target view requires awareness of action category and viewpoint change simultaneously. To address these challenges we propose a novel framework named Pose-guided Action Separable Generative Adversarial Net (PAS-GAN), which utilizes pose to alleviate the difficulty of this task. First, we propose a recurrent pose-transformation module which transforms actions from the source view to the target view and generates novel view pose sequence in 2D coordinate space. Second, a well-transformed pose sequence enables us to separatethe action and background in the target view. We employ a novel local-global spatial transformation module to effectively generate sequential video features in the target view using these action and background features. Finally, the generated video features are used to synthesize human action with the help of a 3D decoder. Moreover, to focus on dynamic action in the video, we propose a novel multi-scale action-separable loss which further improves the video quality. We conduct extensive experiments on two large-scale multi-view human action datasets, NTU-RGBD and PKU-MMD, demonstrating the effectiveness of PAS-GAN which outperforms existing approaches.

****

COCOA: Context-Conditional Adaptation for Recognizing Unseen Classes in Unseen D omains

Puneet Mangla, Shivam Chandhok, Vineeth N Balasubramanian, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 865-874

Recent progress towards designing models that can generalize to unseen domains (i.e domain generalization) or unseen classes (i.e zero-shot learning) has embarked interest towards building models that can tackle both domain-shift and semantic shift simultaneously (i.e zero-shot domain generalization). For models to generalize to unseen classes in unseen domains, it is crucial to learn feature representation that preserves class-level (domain-invariant) as well as domain-specific information. Motivated from the success of generative zero-shot approaches, we propose a feature generative framework integrated with a COntext COnditional Adaptive (COCOA) Batch-Normalization layer to seamlessly integrate class-level semantic and domain-specific information. The generated visual features better capture the underlying data distribution enabling us to generalize to unseen classes and domains at test-time. We thoroughly evaluate our approach on established large-scale benchmarks -- DomainNet, DomainNet-LS (Limited Sources) -- as well as a new CUB-Corruptions benchmark, and demonstrate promising performance over baselines and state-of-the-art methods. We show detailed ablations and analysis to verify that our proposed approach indeed allows us to generate better quality visual features relevant for zero-shot domain generalization.

****

Action Anticipation Using Latent Goal Learning

Debaditya Roy, Basura Fernando; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2745-2753

To get something done, humans perform a sequence of actions dictated by a goal. So, predicting the next action in the sequence becomes easier once we know the goal that guides the entire activity. We present an action anticipation model that uses goal information in an effective manner. Specifically, we use a latent goal representation as a proxy for the "real goal" of the sequence and use this goal information when predicting the next action. We design a model to compute the latent goal representation from the observed video and use it to predict the next action. We also exploit two properties of goals to propose new losses for training the model. First, the effect of the next action should be closer to the latent goal than the observed action, termed as "goal closeness". Second, the latent goal should remain consistent before and after the execution of the next action which we coined as "goal consistency". Using this technique, we obtain state-of-the-art action anticipation performance on scripted datasets 50Salads and Breakfast that have predefined goals in all their videos. We also evaluate the latent goal-based model on EPIC-KITCHENS55 which is an unscripted dataset with multiple goals being pursued simultaneously. Even though this is not an ideal setup for using latent goals, our model is able to predict the next noun better than existing approaches on both seen and unseen kitchens in the test set.

****

DAQ: Channel-Wise Distribution-Aware Quantization for Deep Image Super-Resolution Networks

Cheeun Hong, Heewon Kim, Sungyong Baik, Junghun Oh, Kyoung Mu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2675-2684

Since the resurgence of deep neural networks (DNNs), image super-resolution (SR) has recently seen a huge progress in improving the quality of low resolution images, however at the great cost of computations and resources. Recently, there has been several efforts to make DNNs more efficient via quantization. However, SR demands pixel-level accuracy in the system, it is more difficult to perform quantization without significantly sacrificing SR performance. To this end, we introduce a new ultra-low precision yet effective quantization approach specifically designed for SR. In particular, we observe that in recent SR networks, each channel has different distribution characteristics. Thus we propose a channel-wise distribution-aware quantization scheme. Experimental results demonstrate that our proposed quantization, dubbed Distribution-Aware Quantization (DAQ), manages to greatly reduce the computational and resource costs without the significant sacrifice in SR performance, compared to other quantization methods.

****

Federated Multi-Target Domain Adaptation

Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1424-1433

Federated learning methods enable us to train machine learning models on distributed user data while preserving its privacy. However, it is not always feasible to obtain high-quality supervisory signals from users, especially for vision tasks. Unlike typical federated settings with labeled client data, we consider a more practical scenario where the distributed client data is unlabeled, and a centralized labeled dataset is available on the server. We further take the server-client and inter-client domain shifts into account and pose a domain adaptation problem with one source (centralized server data) and multiple targets (distributed client data). Within this new Federated Multi-Target Domain Adaptation (FMTDA) task, we analyze the model performance of existing domain adaptation methods and propose an effective DualAdapt method to address the new challenges. Extensive experimental results on image classification and semantic segmentation tasks demonstrate that our method achieves high accuracy, incurs minimal communication cost, and requires low computational resources on client devices.

****

Single Source One Shot Reenactment Using Weighted Motion From Paired Feature Points

Soumya Tripathy, Juho Kannala, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2715-2724

Image reenactment is a task where the target object in the source image imitates the motion represented in the driving image. One of the most common reenactment tasks is face image animation. The major challenge in the current face reenactment approaches is to distinguish between facial motion and identity. For this reason, the previous models struggle to produce high-quality animations if the driving and source identities are different (cross-person reenactment). We propose a new (face) reenactment model that learns shape-independent motion features in a self-supervised setup. The motion is represented using a set of paired feature points extracted from the source and driving images simultaneously. The model is generalized to multiple reenactment tasks including faces and non-face objects using only a single source image. The extensive experiments show that the model faithfully transfers the driving motion to the source while retaining the source identity intact.

****

Evaluating the Robustness of Semantic Segmentation for Autonomous Driving Against Real-World Adversarial Patch Attacks

Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, Giorgio Buttazzo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2280-2289

Deep learning and convolutional neural networks allow achieving impressive performance in computer vision tasks, such as object detection and semantic segmentation (SS). However, recent studies have shown evident weaknesses of such models against adversarial perturbations. In a real-world scenario instead, like autonomous driving, more attention should be devoted to real-world adversarial examples (RWAEs), which are physical objects (e.g., billboards and printable patches) optimized to be adversarial to the entire perception pipeline. This paper presents an in-depth evaluation of the robustness of popular SS models by testing the effects of both digital and real-world adversarial patches. These patches are crafted with powerful attacks enriched with a novel loss function. Firstly, an investigation on the Cityscapes dataset is conducted by extending the Expectation Over Transformation (EOT) paradigm to cope with SS. Then, a novel attack optimization, called scene-specific attack, is proposed. Such an attack leverages the CARLA driving simulator to improve the transferability of the proposed EOT-based attack to a real 3D environment. Finally, a printed physical billboard containing a n adversarial patch was tested in an outdoor driving scenario to assess the feasibility of the studied attacks in the real world. Exhaustive experiments revealed that the proposed attack formulations outperform previous work to craft both digital and real-world adversarial patches for SS. At the same time, the experimental results showed how these attacks are notably less effective in the real world, hence questioning the practical relevance of adversarial attacks to SS models for autonomous/assisted driving.

****

Generalized Facial Manipulation Detection With Edge Region Feature Extraction

Dong-Keon Kim, Kwang-Su Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2828-2838

This paper presents a generalized and robust face manipulation detection method based on the edge region features appearing in images. Most contemporary face synthesis processes include color awkwardness reduction but damage the natural fingerprint in the edge region. In addition, these color correction processes do not proceed in the non-face background region. We also observe that the synthesis process does not consider the natural properties of the image appearing in the time domain. Considering these observations, we propose a facial forensic framework that utilizes pixel-level color features appearing in the edge region of the whole image. Furthermore, our framework includes a 3D-CNN classification model that interprets the extracted color features spatially and temporally. Unlike other existing studies, we conduct authenticity determination by considering all features extracted from multiple frames within one video. Through extensive experiments, including real-world scenarios to evaluate generalized detection ability, we show that our framework outperforms state-of-the-art facial manipulation detection technologies in terms of accuracy and robustness.

****

Densely-Packed Object Detection via Hard Negative-Aware Anchor Attention

Sungmin Cho, Jinwook Paeng, Junseok Kwon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2635-2644

In this paper, we propose a novel densely-packed object detection method based on advanced weighted Hausdorff distance (AWHD) and hard negative-aware anchor (HNAA) attention. Densely-packed object detection is more challenging than conventional object detection due to the high object density and small-size objects. To overcome these challenges, the proposed AWHD improves the conventional weighted Hausdorff distance and obtains an accurate center area map. Using the precise center area map, the proposed HNAA attention determines the relative importance of each anchor and imposes a penalty on hard negative anchors. Experimental results demonstrate that our proposed method based on the AWHD and HNAA attention produces accurate densely-packed object detection results and comparably outperforms other state-of-the-art detection methods. The code is available at here.

****

Meta-Meta Classification for One-Shot Learning

Arkabandhu Chowdhury, Dipak Chaudhari, Swarat Chaudhuri, Chris Jermaine; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 177-186

We present a new approach, called meta-meta classification, to learning in small-data settings. In this approach, one uses a large set of learning problems to design an ensemble of learners, where each learner has high bias and low variance and is skilled at solving a specific type of learning problem. The meta-meta classifier learns how to examine a given learning problem and combine the various learners to solve the problem. The meta-meta learning approach is especially suited to solving few-shot learning tasks, as it is easier to learn to classify a new learning problem with little data than it is to apply a learning algorithm to a small data set. We evaluate the approach on a one-shot, one-class-versus-all classification task and show that it is able to outperform traditional meta-learning as well as ensembling approaches.

****

Non-Semantic Evaluation of Image Forensics Tools: Methodology and Database

Quentin Bammey, Tina Nikoukhah, Marina Gardella, Rafael Grompone von Gioi, Migue l Colom, Jean-Michel Morel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3751-3760

We propose a new method to evaluate image forensics tools, that characterizes what image cues are being used by each detector. Our method enables effortless creation of an arbitrarily large dataset of carefully tampered images in which controlled detection cues are present. Starting with raw images, we alter aspects of the image formation pipeline inside a mask, while leaving the rest of the image intact. This does not change the image's interpretation; we thus call such alterations "non-semantic", as they yield no semantic inconsistencies. This method avoids the painful and often biased creation of convincing semantics. All aspects of image formation (noise, CFA, compression pattern and quality, etc.) can vary independently in both the authentic and tampered parts of the image. Alteration of a specific cue enables precise evaluation of the many forgery detectors that rely on this cue, and of the sensitivity of more generic forensic tools to each specific trace of forgery, and can be used to guide the combination of different methods. Based on this methodology, we create a database and conduct an evaluation of the main state-of-the-art image forensics tools, where we characterize the performance of each method with respect to each detection cue. Check qbammey.github.io/trace for the database and code.

****

Weakly Supervised Branch Network With Template Mask for Classifying Masses in 3D Automated Breast Ultrasound

Daekyung Kim, Chang-Mo Nam, Haesol Park, Mijung Jang, Kyong Joon Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3912-3919

Automated breast ultrasound (ABUS) is being rapidly utilized for screening and diagnosing breast cancer. Breast masses, including cancers shown in ABUS scans, often appear as irregular hypoechoic areas that are hard to distinguish from background shadings. We propose a novel branch network architecture incorporating segmentation information of masses in the training process. By providing the spatial attention effect, the branch network boosts the performance of existing neural network classifiers, helping to learn meaningful features around the mass. For the segmentation information, we leverage the existing radiology reports without additional labeling efforts. The reports should include the characteristics of breast masses, such as shape and orientation, and a template mask can be created in a rule-based manner. Experimental results show that the proposed branch network with a template mask significantly improves the performance of existing classifiers.

****

High Dynamic Range Imaging of Dynamic Scenes With Saturation Compensation but Without Explicit Motion Compensation

Haesoo Chung, Nam Ik Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2951-2961

High dynamic range (HDR) imaging is a highly challenging task since a large amount of information is lost due to the limitations of camera sensors. For HDR imaging, some methods capture multiple low dynamic range (LDR) images with altering exposures to aggregate more information. However, these approaches introduce ghosting artifacts when significant inter-frame motions are present. Moreover, although multi-exposure images are given, we have little information in severely over-exposed areas. Most existing methods focus on motion compensation, i.e., alignment of multiple LDR shots to reduce the ghosting artifacts, but they still produce unsatisfying results. These methods also rather overlook the need to restore the saturated areas. In this paper, we generate well-aligned multi-exposure features by reformulating a motion alignment problem into a simple brightness adjustment problem. In addition, we propose a coarse-to-fine merging strategy with explicit saturation compensation. The saturated areas are reconstructed with similar well-exposed content using adaptive contextual attention. We demonstrate that our method outperforms the state-of-the-art methods regarding qualitative and quantitative evaluations.

****

Inpaint2Learn: A Self-Supervised Framework for Affordance Learning

Lingzhi Zhang, Weiyu Du, Shenghao Zhou, Jiancong Wang, Jianbo Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2665-2674

Perceiving affordances -- the opportunities of interaction in a scene, is a fundamental ability of humans. It is an equally important skill for AI agents and robots to better understand and interact with the world. However, labeling affordances in the environment is not a trivial task. To address this issue, we propose a task-agnostic framework, named Inpaint2Learn, that generates affordance labels in a fully automatic manner and opens the door for affordance learning in the wild. To demonstrate its effectiveness, we apply it to three different tasks: human affordance prediction, Location2Object and 6D object pose hallucination. Our experiments and user studies show that our models, trained with the Inpaint2Learn scaffold, are able to generate diverse and visually plausible results in all three scenarios.

****

RGL-NET: A Recurrent Graph Learning Framework for Progressive Part Assembly

Abhinav Narayan, Rajendra Nagar, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 78-87

Autonomous assembly of objects is an essential task in robotics and 3D computer vision. It has been studied extensively in robotics as a problem of motion planning, actuator control and obstacle avoidance. However, the task of developing a generalized framework for assembly robust to structural variants remains relatively unexplored. In this work, we tackle this problem using a recurrent graph learning framework considering inter-part relations and the progressive update of the part pose. Our network can learn more plausible predictions of shape structure by accounting for priorly assembled parts. Compared to the current state-of-the-art, our network yields up to 10% improvement in part accuracy and up to 15% improvement in connectivity accuracy on the PartNet dataset. Moreover, our resulting latent space facilitates exciting applications such as shape recovery from the point-cloud components. We conduct extensive experiments to justify our design choices and demonstrate the effectiveness of the proposed framework.

****

An Experimental Comparison of Multi-View Stereo Approaches on Satellite Images

Alvaro Gómez, Gregory Randall, Gabriele Facciolo, Rafael Grompone von Gioi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 844-853

Different methods can be applied to satellite images to derive an altitude map from a set of images. In this article we evaluate a set of representative methods from different approaches. We consider true multi-view stereo methods as well as pair-wise ones, classic methods and deep learning based ones, methods already in use on satellite images and others that were originally devised for close range imaging and are adapted to satellite imagery. While deep learning (DL) methods have taken over multi-view stereo reconstruction in the last years, this tendency has not fully reached satellite stereo pipelines that still largely rely on pair-wise classic algorithms. For the comparison, we set-up a framework that allows to interface a DL-based stereo method taken from the computer vision literature with a satellite stereo pipeline. For multi-view stereo algorithms we build on a recently proposed framework originally devised to apply Colmap method to satellite images. Methods are compared on several datasets that include sets of images taken within a few days and sets of images taken months apart. Results show that DL methods have, in general, a good generalization power. In particular, the use of the GANet DL method as the matching step in a pair-wise stereo pipeline is promising as it already performs better than the classic counterpart, even without a specific training.

****

Self-Supervised Knowledge Transfer via Loosely Supervised Auxiliary Tasks

Seungbum Hong, Jihun Yoon, Min-Kook Choi, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3318-3327

Knowledge transfer using convolutional neural networks (CNNs) can help efficiently train a CNN with fewer parameters or maximize the generalization performance under limited supervision. To enable a more efficient transfer of pretrained knowledge under relaxed conditions, we propose a simple yet powerful knowledge transfer methodology without any restrictions regarding the network structure or dataset used, namely self-supervised knowledge transfer (SSKT), via loosely supervised auxiliary tasks. For this, we devise a training methodology that transfers previously learned knowledge to the current training process as an auxiliary task for the target task through self-supervision using a soft label. The SSKT is independent of the network structure and dataset, and is trained differently from existing knowledge transfer methods; hence, it has an advantage in that the prior knowledge acquired from various tasks can be naturally transferred during the training process to the target task. Furthermore, it can improve the generalization performance on most datasets through the proposed knowledge transfer between different problem domains from multiple source networks. SSKT outperforms the other transfer learning methods (KD, DML, MAXL) through experiments under various knowledge transfer settings. The source code will be made available to the public

****

Video Salient Object Detection via Contrastive Features and Attention Modules

Yi-Wen Chen, Xiaojie Jin, Xiaohui Shen, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1320-1329

Video salient object detection aims to find the most visually distinct objects in a video. To explore the temporal dependencies, existing methods usually resort to recurrent neural networks or optical flow. However, these approaches require high computational cost, and tend to accumulate inaccuracies over time. In this paper, we propose a network with attention modules to learn contrastive features for video salient object detection without the high computational temporal modeling techniques. We develop a non-local self-attention scheme to capture the global information in the video frame. A co-attention formulation is utilized to combine the low-level and high-level features. We further apply the contrastive learning to improve the feature representations, where foreground region pairs from the same video are pulled together, and foreground-background region pairs are pushed away in the latent space. The intra-frame contrastive loss helps separate the foreground and background features, and the inter-frame contrastive loss improves the temporal consistency. We conduct extensive experiments on several benchmark datasets for video salient object detection and unsupervised video object segmentation, and show that the proposed method requires less computation, and performs favorably against the state-of-the-art approaches.

****

SWAG-V: Explanations for Video Using Superpixels Weighted by Average Gradients

Thomas Hartley, Kirill Sidorov, Christopher Willis, David Marshall; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 604-613

CNN architectures that take videos as an input are often overlooked when it comes to the development of explanation techniques. This is despite their use in critical domains such as surveillance and healthcare. Explanation techniques developed for these networks must take into account the additional temporal domain if they are to be successful. In this paper we introduce SWAG-V, an extension of SWAG for use with networks that take video as an input. By creating superpixels that incorporate individual frames of the input video we are able to create explanations that better locate regions of the input that are important to the networks prediction. We demonstrate using Kinetics-400 with both the C3D and R(2+1)D network architectures that SWAG-V outperforms Grad-CAM, Grad-CAM++ and Saliency Tubes over a range of common metrics such as explanation accuracy and localisation.

****

SIGNAV: Semantically-Informed GPS-Denied Navigation and Mapping in Visually-Degraded Environments

Alex Krasner, Mikhail Sizintsev, Abhinav Rajvanshi, Han-Pang Chiu, Niluthpol Mithun, Kevin Kaighn, Philip Miller, Ryan Villamil, Supun Samarasekera; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2972-2981

Understanding the perceived scene during navigation enables intelligent robot behaviors. Current vision-based semantic SLAM (Simultaneous Localization and Mapping) systems provide these capabilities. However, their performance decreases in visually-degraded environments, that are common places for critical robotic applications, such as search and rescue missions. In this paper, we present SIGNAV, a real-time semantic SLAM system to operate in perceptually-challenging situations. To improve the robustness for navigation in dark environments, SIGNAV leverages a multi-sensor navigation architecture to fuse vision with additional sensing modalities, including an inertial measurement unit (IMU), LiDAR, and wheel odometry. A new 2.5D semantic segmentation method is also developed to combine both images and LiDAR depth maps to generate semantic labels of 3D mapped points in real time. We demonstrate that the navigation accuracy from SIGNAV in a variety of indoor environments under both normal lighting and dark conditions. SIGNAV also provides semantic scene understanding capabilities in visually-degraded environments. We also show the benefits of semantic information to SIGNAV's performance.

****

Strumming to the Beat: Audio-Conditioned Contrastive Video Textures

Medhini Narasimhan, Shiry Ginosar, Andrew Owens, Alexei A. Efros, Trevor Darrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3761-3770

We introduce a non-parametric approach for infinite video texture synthesis using a representation learned via contrastive learning. We take inspiration from Video Textures, which showed that plausible new videos could be generated from a single one by stitching its frames together in a novel yet consistent order. This classic work, however, was constrained by its use of hand-designed distance metrics, limiting its use to simple, repetitive videos. We draw on recent techniques from self-supervised learning to learn this distance metric, allowing us to compare frames in a manner that scales to more challenging dynamics, and to condition on other data, such as audio. We learn representations for video frames and frame-to-frame transition probabilities by fitting a video-specific model trained using contrastive learning. To synthesize a texture, we randomly sample frames with high transition probabilities to generate diverse temporally smooth videos with novel sequences and transitions. The model naturally extends to an audio-conditioned setting without requiring any fine-tuning. Our model outperforms baselines on human perceptual scores, can handle a diverse range of input videos, and can combine semantic and audio-visual cues in order to synthesize videos that synchronize well with an audio signal.

****

SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water

Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, Andreas Zell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2260-2270

Unmanned Aerial Vehicles (UAVs) are of crucial importance in search and rescue missions in maritime environments due to their flexible and fast operation capabilities. Modern computer vision algorithms are of great interest in aiding such missions. However, they are dependent on large amounts of real-case training data from UAVs, which is only available for traffic scenarios on land. Moreover, current object detection and tracking data sets only provide limited environmental information or none at all, neglecting a valuable source of information. Therefore, this paper introduces a large-scaled visual object detection and tracking benchmark (SeaDronesSee) aiming to bridge the gap from land-based vision systems to sea-based ones. We collect and annotate over 54,000 frames with 400,000 instances captured from various altitudes and viewing angles ranging from 5 to 260 meters and 0 to 90 degrees while providing the respective meta information for altitude, viewing angle and other meta data. We evaluate multiple state-of-the-art c omputer vision algorithms on this newly established benchmark serving as baselines. We provide an evaluation server where researchers can upload their prediction and compare their results on a central leaderboard.

****

Detail Preserving Residual Feature Pyramid Modules for Optical Flow

Libo Long, Jochen Lang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2100-2108

Feature pyramids and iterative refinement have recently led to great progress in optical flow estimation. However, downsampling in feature pyramids can cause blending of foreground objects with the background, which will mislead subsequent decisions in the iterative processing. The results are missing details especially in the flow of thin and of small structures. We propose a novel Residual Feature Pyramid Module (RFPM) which retains important details in the feature map without changing the overall iterative refinement design of the optical flow estimation. RFPM incorporates a residual structure between multiple feature pyramids into a downsampling module that corrects the blending of objects across boundaries. We demonstrate how to integrate our module with two state-of-the-art iterative refinement architectures. Results show that our RFPM visibly reduces flow errors and improves state-of-art performance in the clean pass of Sintel, and is one of the top-performing methods in KITTI. According to the particular modular structure of RFPM, we introduce a special transfer learning approach that can dramatically decrease the training time compared to a typical full optical flow training schedule on multiple datasets.

****

Novel-View Synthesis of Human Tourist Photos

Jonathan Freer, Kwang Moo Yi, Wei Jiang, Jongwon Choi, Hyung Jin Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3069-3076

We present a novel framework for performing novel-view synthesis on human tourist photos. Given a tourist photo from a known scene, we reconstruct the photo in 3D space through modeling the human and the background independently. We generate a deep buffer from a novel view point of the reconstruction and utilize a deep network to translate the buffer into a photo realistic rendering of the novel view. We additionally present a method to relight the renderings, allowing for relighting of both human and background to match either the provided input image or any other. The key contributions of our paper are: 1) a framework for performing novel view synthesis on human tourist photos, 2) an appearance transfer method for relighting of humans to match synthesized backgrounds, and 3) a method for estimating lighting properties from a single human photo.

****

Stylizing 3D Scene via Implicit Representation and HyperNetwork

Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, Wei-Chen Chiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1475-1484

In this work, we aim to address the 3D scene stylization problem - generating stylized images of the scene at arbitrary novel view angles. A straightforward solution is to combine existing novel view synthesis and image/video style transfer approaches, which often leads to blurry results or inconsistent appearance. Inspired by the high quality results of the neural radiance fields (NeRF) method, we propose a joint framework to directly render novel views with the desired style. Our framework consists of two components: an implicit representation of the 3D scene with the neural radiance field model, and a hypernetwork to transfer the style information into the scene representation. To alleviate the training difficulties and memory burden, we propose a two-stage training procedure and a patch sub-sampling approach to optimize the style and content losses with the neural radiance field model. After optimization, our model is able to render consistent novel views at arbitrary view angles with arbitrary style. Both quantitative evaluation and human subject study have demonstrated that the proposed method generates faithful stylization results with consistent appearance across different views.

****

Novel Ensemble Diversification Methods for Open-Set Scenarios
Miriam Farber, Roman Goldenberg, George Leifman, Gal Novich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1485-1494

We revisit existing ensemble diversification approaches and present two novel diversification methods tailored for open-set scenarios. The first method uses a new loss, designed to encourage models disagreement on outliers only, thus alleviating the intrinsic accuracy-diversity trade-off. The second method achieves diversity via automated feature engineering, by training each model to disregard input features learned by previously trained ensemble models. We conduct an extensive evaluation and analysis of the proposed techniques on seven datasets that cover image classification, re-identification and recognition domains. We compare to and demonstrate accuracy improvements over the existing state-of-the-art ensemble diversification methods.

****

Learning With Label Noise for Image Retrieval by Selecting Interactions
Sarah Ibrahimi, Arnaud Sors, Rafael Sampaio de Rezende, Stéphane Clinchant; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2181-2190

Learning with noisy labels is an active research area for image classification. However, the effect of noisy labels on image retrieval has been less studied. In this work, we propose a noise-resistant method for image retrieval named Teacher-based Selection of Interactions, T-SINT, which identifies noisy interactions, i.e. elements in the distance matrix, and selects correct positive and negative interactions to be considered in the retrieval loss by using a teacher-based training setup which contributes to the stability. As a result, it consistently outperforms state-of-the-art methods on high noise rates across benchmark datasets with synthetic noise and more realistic noise.

****

LEAD: Self-Supervised Landmark Estimation by Aligning Distributions of Feature Similarity
Tejan Karmali, Abhinav Atrishi, Sai Sree Harsha, Susmit Agrawal, Varun Jampani, R. Venkatesh Babu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 623-632

In this work, we introduce LEAD, an approach to discover landmarks from an unannotated collection of category-specific images. Existing works in self-supervised landmark detection are based on learning dense (pixel-level) feature representations from an image, which are further used to learn landmarks in a semi-supervised manner. While there have been advances in self-supervised learning of image features for instance-level tasks like classification, these methods do not ensure dense equivariant representations. The property of equivariance is of interest for dense prediction tasks like landmark estimation. In this work, we introduce an approach to enhance the learning of dense equivariant representations in a self-supervised fashion. We follow a two-stage training approach: first, we train a network using the BYOL objective which operates at an instance level. The correspondences obtained through this network are further used to train a dense and compact representation of the image using a lightweight network. We show that having such a prior in the feature extractor helps in landmark detection, even under drastically limited number of annotations while also improving generalization across scale variations.

****

Contrast To Divide: Self-Supervised Pre-Training for Learning With Noisy Labels
Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M. Bronstein, Or Litany; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1657-1667

The success of learning with noisy labels (LNL) methods relies heavily on the success of a warm-up stage where standard supervised training is performed using the full (noisy) training set. In this paper, we identify a "warm-up obstacle": the inability of standard warm-up stages to train high quality feature extractors and avert memorization of noisy labels. We propose "Contrast to Divide" (C2D), a simple framework that solves this problem by pre-training the feature extractor in a self-supervised fashion. Using self-supervised pre-training boosts the performance of existing LNL approaches by drastically reducing the warm-up stage's susceptibility to noise level, shortening its duration, and improving extracted feature quality. C2D works out of the box with existing methods and demonstrates markedly improved performance, especially in the high noise regime, where we get a boost of more than 27% for CIFAR-100 with 90% noise over the previous state of the art. In real-life noise settings, C2D trained on mini-WebVision outperforms previous works both in WebVision and ImageNet validation sets by 3% top-1 accuracy. We perform an in-depth analysis of the framework, including investigating the performance of different pre-training approaches and estimating the effective upper bound of the LNL performance with semi-supervised learning. Code for reproducing our experiments is available at https://github.com/ContrastToDivide/C2D

****

Co-Net: A Collaborative Region-Contour-Driven Network for Fine-to-Finer Medical Image Segmentation

Anran Liu, Xiangsheng Huang, Tong Li, Pengcheng Ma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1046-1055

In this paper, a fine-to-finer segmentation task is investigated driven by region and contour features collaboratively on Glomerular Electron-Dense Deposits (GEDD) in view of the complementary nature of these two types of features. To this end, a novel network (Co-Net) is presented to dynamically use fine saliency segmentation to guide finer segmentation on boundaries. The whole architecture contains double mutually boosted decoders sharing one common encoder. Specifically, a new structure named Global-guided Interaction Module (GIM) is designed to effectively control the information flow and reduce redundancy in the cross-level feature fusion process. At the same time, the global features are used in it to make the features of each layer gain access to richer context, and a fine segmentation map is obtained initially; Discontinuous Boundary Supervision (DBS) strategy is applied to pay more attention to discontinuity positions and modifying segmentation errors on boundaries. At last, Selective Kernel (SK) is used for dynamical aggregation of the region and contour features to obtain a finer segmentation. Our proposed approach is evaluated on an independent GEDD dataset labeled by pathologists and also on open polyp datasets to test the generalization. Ablation studies show the effectiveness of different modules. On all datasets, our proposal achieves high segmentation accuracy and surpasses previous methods.

****

Dataset Knowledge Transfer for Class-Incremental Learning Without Memory

Habib Slim, Eden Belouadah, Adrian Popescu, Darian Onchis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 483-492

Incremental learning enables artificial agents to learn from sequential data. While important progress was made by exploiting deep neural networks, incremental learning remains very challenging. This is particularly the case when no memory of past data is allowed and catastrophic forgetting has a strong negative effect. We tackle class-incremental learning without memory by adapting prediction bias correction, a method which makes predictions of past and new classes more comparable. It was proposed when a memory is allowed and cannot be directly used without memory, since samples of past classes are required. We introduce a two-step learning process which allows the transfer of bias correction parameters between reference and target datasets. Bias correction is first optimized offline on reference datasets which have an associated validation memory. The obtained correction parameters are then transferred to target datasets, for which no memory is available. The second contribution is to introduce a finer modeling of bias correction by learning its parameters per incremental state instead of the usual past vs. new class modeling. The proposed dataset knowledge transfer is applicable to any incremental method which works without memory. We test its effectiveness by applying it to four existing methods. Evaluation with four target datasets a

nd different configurations shows consistent improvement, with practically no computational and memory overhead.
****

Enhanced Correlation Matching Based Video Frame Interpolation
Sungho Lee, Narae Choi, Woong Il Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2839-2847
We propose a novel DNN based framework called the Enhanced Correlation Matching based Video Frame Interpolation Network to support high resolution like 4K, which has a large scale of motion and occlusion. Considering the extensibility of the network model according to resolution, the proposed scheme employs the recurrent pyramid architecture that shares the parameters among each pyramid layer for the optical flow estimation. In the proposed flow estimation, the optical flows are recursively refined by tracing the location with maximum correlation. The forward warping based correlation matching enables to improve the accuracy of flow update by excluding incorrectly warped features around the occlusion area. Based on the final bi-directional flows, the intermediate frame at arbitrary temporal position is synthesized using the warping and blending network and it is further improved by refinement network. Experiment results demonstrate that the proposed scheme outperforms the previous works at 4K video data and low-resolution benchmark datasets as well in terms of objective and subjective quality with the smallest number of model parameters.
****

Non-Local Attention Improves Description Generation for Retinal Images
Jia-Hong Huang, Ting-Wei Wu, C.-H. Huck Yang, Zenglin Shi, I-Hung Lin, Jesper Tegner, Marcel Worring; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1606-1615
Automatically generating medical reports from retinal images is a difficult task in which an algorithm must generate semantically coherent descriptions for a given retinal image. Existing methods mainly rely on the input image to generate descriptions. However, many abstract medical concepts or descriptions cannot be generated based on image information only. In this work, we integrate additional information to help solve this task; we observe that early in the diagnosis process, ophthalmologists have usually written down a small set of keywords denoting important information. These keywords are then subsequently used to aid the later creation of medical reports for a patient. Since these keywords commonly exist and are useful for generating medical reports, we incorporate them into automatic report generation. Since we have two types of inputs - expert-defined unordered keywords and images - effectively fusing features from these different modalities is challenging. To that end, we propose a new keyword-driven medical report generation method based on a non-local attention-based multi-modal feature fusion approach, TransFuser, which is capable of fusing features from different types of inputs based on such attention. Our experiments show the proposed method successfully captures the mutual information of keywords and image content. We further show our proposed keyword-driven generation model reinforced by the TransFuser is superior to baselines under the popular text evaluation metrics BLEU, CIDEr, and ROUGE. TransFuser Github:https://github.com/Jhhuangkay/Non-local-Attention-Improves-Description-Generation-for-Retinal-Images.
****

Style Agnostic 3D Reconstruction via Adversarial Style Transfer
Felix Petersen, Bastian Goldluecke, Oliver Deussen, Hilde Kuehne; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3664-3673
Reconstructing the 3D geometry of an object from an image is a major challenge in computer vision. Recently introduced differentiable renderers can be leveraged to learn the 3D geometry of objects from 2D images, but those approaches require additional supervision to enable the renderer to produce an output that can be compared to the input image. This can be scene information or constraints such as object silhouettes, uniform backgrounds, material, texture, and lighting. In this paper, we propose an approach that enables a differentiable rendering-based learning of 3D objects from images with backgrounds without the need for silhou

ette supervision. Instead of trying to render an image close to the input, we propose an adversarial style-transfer and domain adaptation pipeline that allows to translate the input image domain to the rendered image domain. This allows us to directly compare between a translated image and the differentiable rendering of a 3D object reconstruction in order to train the 3D object reconstruction network. We show that the approach learns 3D geometry from images with backgrounds and provides a better performance than constrained methods for single-view 3D object reconstruction on this task.
****
A Deep Insight Into Measuring Face Image Utility With General and Face-Specific Image Quality Metrics
Biying Fu, Cong Chen, Olaf Henniger, Naser Damer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 905-914
Quality scores provide a measure to evaluate the utility of biometric samples for biometric recognition. Biometric recognition systems require high-quality samples to achieve optimal performance. This paper focuses on face images and the measurement of face image utility with general and face-specific image quality metrics. While face-specific metrics rely on features of aligned face images, general image quality metrics can be used on the global image and relate to human perceptions. In this paper, we analyze the gap between the general image quality metrics and the face image quality metrics. Our contribution lies in a thorough examination of how different the image quality assessment algorithms relate to the utility for the face recognition task. The results of image quality assessment algorithms are further compared with those of dedicated face image quality assessment algorithms. In total, 25 different quality metrics are evaluated on three face image databases, BioSecure, LFW, and VGGFace2 using three open-source face recognition solutions, SphereFace, ArcFace, and FaceNet. Our results reveal a clear correlation between learned image metrics to face image utility even without being specifically trained as a face utility measure. Individual handcrafted features lack general stability and perform significantly worse than general face-specific quality metrics. We additionally provide a visual insight into the image areas contributing to the quality score of a selected set of quality assessment methods.
****
AirCamRTM: Enhancing Vehicle Detection for Efficient Aerial Camera-Based Road Traffic Monitoring
Rafael Makrigiorgis, Nicolas Hadjittoouli, Christos Kyrkou, Theocharis Theocharides; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2119-2128
Efficient road traffic monitoring is playing a fundamental role in successfully resolving traffic congestion in cities.Unmanned Aerial Vehicles (UAVs) or drones equipped with cameras are an attractive proposition to provide flexible and infrastructure-free traffic monitoring. However, real-time traffic monitoring from UAV imagery poses several challenges, due to the large image sizes and presence of non relevant targets. In this paper, we propose the AirCam-RTM framework that combines road segmentation and vehicle detection to focus only on relevant vehicles, which as a result, improves the monitoring performance by approximately 2x and provides approximately 18% accuracy improvement. Furthermore,through a real experimental setup we qualitatively evaluate the performance of the proposed approach, and also demonstrate how it can be used for real-time traffic monitoring and management using UAVs.
****
SEGA: Semantic Guided Attention on Visual Prototype for Few-Shot Learning
Fengyuan Yang, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1056-1066
Teaching machines to recognize a new category based on few training samples especially only one remains challenging owing to the incomprehensive understanding of the novel category caused by the lack of data. However, human can learn new classes quickly even given few samples since human can tell what discriminative features should be focused on about each category based on both the visual and sem

antic prior knowledge. To better utilize those prior knowledge, we propose the SEmantic Guided Attention (SEGA) mechanism where the semantic knowledge is used to guide the visual perception in a top-down manner about what visual features should be paid attention to when distinguishing a category from the others. As a result, the embedding of the novel class even with few samples can be more discriminative. Concretely, a feature extractor is trained to embed few images of each novel class into a visual prototype with the help of transferring visual prior knowledge from base classes. Then we learn a network that maps semantic knowledge to category-specific attention vectors which will be used to perform feature selection to enhance the visual prototypes. Extensive experiments on miniImageNet, tieredImageNet, CIFAR-FS, and CUB indicate that our semantic guided attention realizes anticipated function and outperforms state-of-the-art results.

****

Sign Language Translation With Hierarchical Spatio-Temporal Graph Neural Network

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, Zhiyong Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3367-3376

Sign language translation (SLT), which generates text in a spoken language from visual content in a sign language, is important to assist the hard-of-hearing community for their communications. Inspired by neural machine translation (NMT), most existing SLT studies adopt a general sequence to sequence learning strategy. However, SLT is significantly different from conventional NMT tasks since sign languages convey messages through multiple aspects simultaneously such as hand poses, relative positions and body movements. Therefore, in this paper, the unique characteristics of the signing poses of sign languages is utilized to formulate hierarchical spatio-temporal graph representations of signing poses, including both high-level and fine-level graphs of which each vertex characterizes a specified body part and the edges represent the interactions between any two vertices. Specifically, high-level graphs represent the interactions between key regions such as hands and face, and fine-level graphs represent relationships between the joints of each hand and landmarks of facial regions. To this end, a novel deep learning architecture, namely hierarchical spatio-temporal graph neural network (HST-GNN), is proposed to learn such graph representations. In addition, graph convolutions and graph self-attentions with neighborhood context are proposed to characterize both the local and the global graph properties. Experimental results on benchmark datasets demonstrated the the performance.

****

Pose and Joint-Aware Action Recognition

Anshul Shah, Shlok Mishra, Ankan Bansal, Jun-Cheng Chen, Rama Chellappa, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3850-3860

Recent progress on action recognition has mainly focused on RGB and optical flow features. In this paper, we approach the problem of joint-based action recognition. Unlike other modalities, constellation of joints and their motion generate models with succinct human motion information for activity recognition. We present a new model for joint-based action recognition, which first extracts motion features from each joint separately through a shared motion encoder before performing collective reasoning. Our joint selector module re-weights the joint information to select the most discriminative joints for the task. We also propose a novel joint-contrastive loss that pulls together groups of joint features which convey the same action. We strengthen the joint-based representations by using a geometry-aware data augmentation technique which jitters pose heatmaps while retaining the dynamics of the action. We show large improvements over the current state-of-the-art joint-based approaches on JHMDB, HMDB, Charades, AVA action recognition datasets. A late fusion with RGB and Flow-based approaches yields additional improvements. Our model also outperforms the existing baseline on Mimetics, a dataset with out-of-context actions.

****

QUALIFIER: Question-Guided Self-Attentive Multimodal Fusion Network for Audio Visual Scene-Aware Dialog

Muchao Ye, Quanzeng You, Fenglong Ma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 248-256

Audio video scene-aware dialog (AVSD) is a new but more challenging visual question answering (VQA) task because of the higher complexity of feature extraction and fusion brought by the additional modalities. Although recent methods have achieved early success in improving feature extraction technique for AVSD, the technique of feature fusion still needs further investigation. In this paper, inspired by the success of self-attention mechanism and the importance of understanding questions for VQA answering, we propose a question-guided self-attentive multi-modal fusion network (QUALIFIER) (QUALIFIER) to improve the AVSD practice in the stage of feature fusion and answer generation. Specifically, after extracting features and learning a comprehensive feature for each modality, we first use the designed self-attentive multi-modal fusion (SMF) module to aggregate each feature with the correlated information learned from others. Later, by prioritizing the question feature, we concatenate it with each fused feature to guide the generation of a natural language response to the question. As for experimental results, QUALIFIER shows better performance than other baseline methods in the large-scale AVSD dataset named DSTC7. Additionally, the human evaluation and ablation study results also demonstrate the effectiveness of our network architecture.

****

Domain Generalization Through Audio-Visual Relative Norm Alignment in First Person Action Recognition

Mirco Planamente, Chiara Plizzari, Emanuele Alberti, Barbara Caputo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1807-1818

First person action recognition is becoming an increasingly researched area thanks to the rising popularity of wearable cameras. This is bringing to light cross-domain issues that are yet to be addressed in this context. Indeed, the information extracted from learned representations suffers from an intrinsic "environmental bias". This strongly affects the ability to generalize to unseen scenarios, limiting the application of current methods to real settings where labeled data are not available during training. In this work, we introduce the first domain generalization approach for egocentric activity recognition, by proposing a new audio-visual loss, called Relative Norm Alignment loss. It re-balances the contributions from the two modalities during training, over different domains, by aligning their feature norm representations. Our approach leads to strong results in domain generalization on both EPIC-Kitchens-55 and EPIC-Kitchens-100, as demonstrated by extensive experiments, and can be extended to work also on domain adaptation settings with competitive results.

****

Detection and Localization of Facial Expression Manipulations

Ghazal Mazaheri, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1035-1045

Concerns regarding the wide-spread use of forged images and videos in social media necessitate precise detection of such fraud. Facial manipulations can be created by Identity swap (DeepFake) or Expression swap. Contrary to the identity swap, which can easily be detected with novel deepfake detection methods, expression swap detection has not yet been addressed extensively. The importance of facial expressions in inter-person communication is known. Consequently, it is important to develop methods that can detect and localize manipulations in facial expressions. To this end, we present a novel framework to exploit the underlying feature representations of facial expressions learned from expression recognition models to identify the manipulated features. Using discriminative feature maps extracted from a facial expression recognition framework, our manipulation detector is able to localize the manipulated regions of input images and videos. On the Face2Face dataset, (abundant expression manipulation), and NeuralTextures dataset (facial expressions manipulation corresponding to the mouth regions), our method achieves higher accuracy for both classification and localization of manipulations compared to state-of-the-art methods. Furthermore, we demonstrate that our method performs at-par with the state-of-the-art methods in cases where the ex pression is not manipulated, but rather the identity is changed, leading to a generalized approach for facial manipulation detection.

****

Unsupervised Learning for Human Sensing Using Radio Signals

Tianhong Li, Lijie Fan, Yuan Yuan, Dina Katabi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3288-3297

There is a growing literature demonstrating the feasibility of using Radio Frequency (RF) signals to enable key computer vision tasks in the presence of occlusions and poor lighting. It leverages that RF signals traverse walls and occlusions to deliver through-wall pose estimation, action recognition, scene captioning, and human re-identification. However, unlike RGB datasets which can be labeled by human workers, labeling RF signals is a daunting task because such signals are not human interpretable. Yet, it is fairly easy to collect unlabelled RF signals. It would be highly beneficial to use such unlabeled RF data to learn useful representations in an unsupervised manner. Thus, in this paper, we explore the feasibility of adapting RGB-based unsupervised representation learning to RF signals. We show that while contrastive learning has emerged as the main technique for unsupervised representation learning from images and videos, such methods produce poor performance when applied to sensing humans using RF signals. In contrast, predictive unsupervised learning methods learn high-quality representations that can be used for multiple downstream RF-based sensing tasks. Our empirical results show that this approach outperforms state-of-the-art RF-based human sensing on various tasks, opening the possibility of unsupervised representation learning from this novel modality.

****

Joint Classification and Trajectory Regression of Online Handwriting Using a Multi-Task Learning Approach

Felix Ott, David Rügamer, Lucas Heublein, Bernd Bischl, Christopher Mutschler; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 266-276

Multivariate Time Series (MTS) classification is important in various applications such as signature verification, person identification, and motion recognition. In deep learning these classification tasks are usually learned using the cross-entropy loss. A related yet different task is predicting trajectories observed as MTS. Important use cases include handwriting reconstruction, shape analysis, and human pose estimation. The goal is to align an arbitrary dimensional time series with its ground truth as accurately as possible while reducing the error in the prediction with a distance loss and the variance with a similarity loss. Although learning both losses with Multi-Task Learning (MTL) helps to improve trajectory alignment, learning often remains difficult as both tasks are contradictory. We propose a novel neural network architecture for MTL that notably improves the MTS classification and trajectory regression performance in online handwriting (OnHW) recognition. We achieve this by jointly learning the cross-entropy loss in combination with distance and similarity losses. On an OnHW task of handwritten characters with multivariate inertial and visual data inputs we are able to achieve crucial improvements (lower error with less variance) of trajectory prediction while still improving the character classification accuracy in comparison to models trained on the individual tasks.

****

An Investigation of Critical Issues in Bias Mitigation Techniques

Robik Shrestha, Kushal Kafle, Christopher Kanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1943-1954

A critical problem in deep learning is that systems learn inappropriate biases, resulting in their inability to perform well on minority groups. This has led to the creation of multiple algorithms that endeavor to mitigate bias. However, it is not clear how effective these methods are. This is because study protocols differ among papers, systems are tested on datasets that fail to test many forms of bias, and systems have access to hidden knowledge or are tuned specifically to the test set. To address this, we introduce an improved evaluation protocol, sensible metrics, and a new dataset, which enables us to ask and answer critical questions about bias mitigation algorithms. We evaluate seven state-of-the-art algorithms using the same network architecture and hyperparameter selection policy across three benchmark datasets. We introduce a new dataset called BiasedMNIST that enables the assessment of robustness to multiple bias sources. We use BiasedMNIST and a visual question answering (VQA) benchmark to assess robustness to hidden biases. Rather than only tuning to the test set distribution, we study robustness across different tuning distributions, which is critical because for many applications the test distribution may not be known during development. We find that algorithms exploit hidden biases, are unable to scale to multiple forms of bias, and are highly sensitive to the choice of tuning set. Based on our findings, we implore the community to adopt more rigorous assessment of future bias mitigation methods. All data, code and results will be made publicly available.

****

Dynamic CNNs Using Uncertainty To Overcome Domain Generalization for Surgical Instrument Localization

Markus Philipp, Anna Alperovich, Marielena Gutt-Will, Andrea Mathis, Stefan Saur, Andreas Raabe, Franziska Mathis-Ullrich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3612-3621

Due to the limited amount of available annotated data in the medical field, domain generalization for applications in computer-assisted surgery is essential. Our work addresses this problem for the task of surgical instrument tip localization in neurosurgery, which is a classical step towards computer-assisted surgery. We propose an uncertainty-based CNN approach that dynamically selects the most relevant data source by incorporating its own uncertainty into the inference. In addition, the estimated uncertainty can visualize and easily explain the network's decision. Quantitative and qualitative evaluations show that our method outperforms state of the art approaches for large domain shifts and results are on-par for in-domain applications. Further increasing domain shifts by testing on different surgical disciplines, eye and laparoscopic surgeries, proves the generalization capabilities of the proposed method.

****

Cleaning Noisy Labels by Negative Ensemble Learning for Source-Free Unsupervised Domain Adaptation

Waqar Ahmed, Pietro Morerio, Vittorio Murino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1616-1625

Conventional Unsupervised Domain Adaptation (UDA) methods presume source and target domain data to be simultaneously available during training. Such an assumption may not hold in practice, as source data is often inaccessible (e.g., due to privacy reasons). On the contrary, a pre-trained source model is usually available, which performs poorly on target due to the well-known domain shift problem. This translates into a significant amount of misclassifications, which can be interpreted as structured noise affecting the inferred target pseudo-labels. In this work, we cast UDA as a pseudo-label refinery problem in the challenging source-free scenario. We propose Negative Ensemble Learning (NEL) technique, a unified method for adaptive noise filtering and progressive pseudo-label refinement. NEL is devised to tackle noisy pseudo-labels by enhancing diversity in ensemble members with different stochastic (i) input augmentation and (ii) feedback. The latter is achieved by leveraging the novel concept of Disjoint Residual Labels, which allow propagating diverse information to the different members. Eventually, a single model is trained with the refined pseudo-labels, which leads to a robust performance on the target domain. Extensive experiments show that the proposed method achieves state-of-the-art performance on major UDA benchmarks, such as Digit5, PACS, Visda-C, and DomainNet, without using source data samples at all.

****

FLUID: Few-Shot Self-Supervised Image Deraining

Shyam Nandan Rai, Rohit Saluja, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3077-3086

Self-supervised methods have shown promising results in denoising and dehazing tasks, where the collection of the paired dataset is challenging and expensive. H owever, we find that these methods fail to remove the rain streaks when applied for image deraining tasks. The method's poor performance is due to the explicit assumptions: (i) the distribution of noise or haze is uniform and (ii) the value of a noisy or hazy pixel is independent of its neighbors. The rainy pixels are non-uniformly distributed, and it is not necessarily dependant on its neighboring pixels. Hence, we conclude that the self-supervised method needs to have some prior knowledge about rain distribution to perform the deraining task. To provide this knowledge, we hypothesize a network trained with minimal supervision to estimate the likelihood of rainy pixels. This leads us to our proposed method called FLUID: Few Shot Self-Supervised Image Deraining. We perform extensive experiments and comparisons with existing image deraining and few-shot image-to-image translation methods on Rain 100L and DDN-SIRR datasets containing real and synthetic rainy images. In addition, we use the Rainy Cityscapes dataset to show that our method trained in a few-shot setting can improve semantic segmentation and object detection in rainy conditions. Our approach obtains a mIoU gain of 51.20 over the current best-performing deraining method.

****

SAC: Semantic Attention Composition for Text-Conditioned Image Retrieval

Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4021-4030

The ability to efficiently search for images is essential for improving the user experiences across various products. Incorporating user feedback, via multi-modal inputs, to navigate visual search can help tailor retrieved results to specific user queries. We focus on the task of text-conditioned image retrieval that utilizes support text feedback alongside a reference image to retrieve images that concurrently satisfy constraints imposed by both inputs. The task is challenging since it requires learning composite image-text features by incorporating multiple cross-granular semantic edits from text feedback and then applying the same to visual features. To address this, we propose a novel framework SAC which resolves the above in two major steps: "where to see" (Semantic Feature Attention) and "how to change" (Semantic Feature Modification). We systematically show how our architecture streamlines the generation of text-aware image features by removing the need for various modules required by other state-of-art techniques. We present extensive quantitative, qualitative analysis, and ablation studies, to show that our architecture SAC outperforms existing techniques by achieving state-of-the-art performance on 3 benchmark datasets: FashionIQ, Shoes, and Birds-to-Words while supporting natural language feedback of varying lengths.

****

Robust 3D Garment Digitization From Monocular 2D Images for 3D Virtual Try-On Systems

Sahib Majithia, Sandeep N. Parameswaran, Sadbhavana Babar, Vikram Garg, Astitva Srivastava, Avinash Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3428-3438

In this paper, we develop a robust 3D garment digitization solution that can generalize well on real-world fashion catalog images with cloth texture occlusions and large body pose variations. We assumed fixed topology parametric template mesh models for known types of garments (e.g., T-shirts, Trousers) and perform mapping of high-quality texture from an input catalog image to UV map panels corresponding to the parametric mesh model of the garment. We achieve this by first predicting a sparse set of 2D landmarks on the boundary of the garments. Subsequently, we use these landmarks to perform Thin-Plate-Spline-based texture transfer on UV map panels. Subsequently, we employ a deep texture inpainting network to fill the large holes (due to view variations & self-occlusions) in TPS output to generate consistent UV maps. Furthermore, to train the supervised deep networks for landmark prediction & texture inpainting tasks, we generated a large set of synthetic data with varying texture and lighting imaged from various views with the human present in a wide variety of poses. Additionally, we manually annotated a small set of fashion catalog images crawled from online fashion e-commerce platforms to finetune. We conduct thorough empirical evaluations and show impress ive qualitative results of our proposed 3D garment texture solution on fashion catalog images. Such 3D garment digitization helps us solve the challenging task of enabling 3D Virtual Try-on.

****

Contextual Proposal Network for Action Localization

He-Yen Hsieh, Ding-Jie Chen, Tyng-Luh Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2129-2138

This paper investigates the problem of Temporal Action Proposal (TAP) generation, which aims to provide a set of high-quality video segments that potentially contain actions events locating in long untrimmed videos. Based on the goal to distill available contextual information, we introduce a Contextual Proposal Network (CPN) composing of two context-aware mechanisms. The first mechanism, i.e., feature enhancing, integrates the inception-like module with long-range attention to capture the multi-scale temporal contexts for yielding a robust video segment representation. The second mechanism, i.e., boundary scoring, employs the bi-directional recurrent neural networks (RNN) to capture bi-directional temporal contexts that explicitly model actionness, background, and confidence of proposals. While generating and scoring proposals, such bi-directional temporal contexts are helpful to retrieve high-quality proposals of low false positives for covering the video action instances. We conduct experiments on two challenging datasets of ActivityNet-1.3 and THUMOS-14 to demonstrate the effectiveness of the proposed Contextual Proposal Network (CPN). In particular, our method respectively surpasses state-of-the-art TAP methods by 1.54% AUC on ActivityNet-1.3 test split and by 0.61% AR@200 on THUMOS-14 dataset.

****

Tensor Feature Hallucination for Few-Shot Learning

Michalis Lazarou, Tania Stathaki, Yannis Avrithis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3500-3510

Few-shot learning addresses the challenge of learning how to address novel tasks given not just limited supervision but limited data as well. An attractive solution is synthetic data generation. However, most such methods are overly sophisticated, focusing on high-quality, realistic data in the input space. It is unclear whether adapting them to the few-shot regime and using them for the downstream task of classification is the right approach. Previous works on synthetic data generation for few-shot classification focus on exploiting complex models, e.g. a Wasserstein GAN with multiple regularizers or a network that transfers latent diversities from known to novel classes. We follow a different approach and investigate how a simple and straightforward synthetic data generation method can be used effectively. We make two contributions, namely we show that: (1) using a simple loss function is more than enough for training a feature generator in the few-shot setting; and (2) learning to generate tensor features instead of vector features is superior. Extensive experiments on miniImagenet, CUB and CIFAR-FS datasets show that our method sets a new state of the art, outperforming more sophisticated few-shot data augmentation methods. The source code can be found at https://github.com/MichalisLazarou/TFH_fewshot.

****

Semi-Supervised Multi-Task Learning for Semantics and Depth

Yufeng Wang, Yi-Hsuan Tsai, Wei-Chih Hung, Wenrui Ding, Shuo Liu, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2505-2514

Multi-Task Learning (MTL) aims to enhance the model generalization by sharing representations between related tasks for better performance. Typical MTL methods are jointly trained with the complete multitude of ground-truths for all tasks simultaneously. However, one single dataset may not contain the annotations for each task of interest. To address this issue, we propose the Semi-supervised Multi-Task Learning (SemiMTL) method to leverage the available supervisory signals from different datasets, particularly for semantic segmentation and depth estimation tasks. To this end, we design an adversarial learning scheme in our semi-supervised training by leveraging unlabeled data to optimize all the task branches simultaneously and accomplish all tasks across datasets with partial annotations

. We further present a domain-aware discriminator structure with various alignment formulations to mitigate the domain discrepancy issue among datasets. Finally, we demonstrate the effectiveness of the proposed method to learn across different datasets on challenging street view and remote sensing benchmarks.

****

Extracting Vignetting and Grain Filter Effects From Photos

Abdelrahman Abdelhamed, Jonghwa Yim, Abhijith Punnappurath, Michael S. Brown, Jihwan Choe, Kihwan Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1191-1199

Most smartphones support the use of real-time camera filters to impart visual effects to captured images. Currently, such filters come preinstalled on-device or need to be downloaded and installed before use (e.g., Instagram filters). Recent work [24] proposed a method to extract a camera filter directly from an example photo that has already had a filter applied. The work in [24] focused only on the color and tonal aspects of the underlying filter. In this paper, we introduce a method to extract two spatially varying effects commonly used by on-device camera filters---namely, image vignetting and image grain. Specifically, we show how to extract the parameters for vignetting and image grain present in an example image and replicate these effects as an on-device filter. We use lightweight CNNs to estimate the filter parameters and employ efficient techniques---isotropic Gaussian filters and simplex noise---for regenerating the filters. Our design achieves a reasonable trade-off between efficiency and realism. We show that our method can extract vignetting and image grain filters from stylized photos and replicate the filters on captured images more faithfully, as compared to color and style transfer methods. Our method is significantly efficient and has been already deployed to millions of flagship smartphones.

****

Unsupervised Robust Domain Adaptation Without Source Data

Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2009-2018

We study the problem of robust domain adaptation in the context of unavailable target labels and source data. The considered robustness is against adversarial perturbations. This paper aims at answering the question of finding the right strategy to make the target model robust and accurate in the setting of unsupervised domain adaptation without source data. The major findings of this paper are: (i) robust source models can be transferred robustly to the target; (ii) robust domain adaptation can greatly benefit from non-robust pseudo-labels and the pair-wise contrastive loss. The proposed method of using non-robust pseudo-labels performs surprisingly well on both clean and adversarial samples, for the task of image classification. We show a consistent performance improvement of over 10% in accuracy against the tested baselines on four benchmark datasets.

****

Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection

Su Pang, Daniel Morris, Hayder Radha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 187-196

When compared to single modality approaches, fusion-based object detection methods often require more complex models to integrate heterogeneous sensor data, and use more GPU memory and computational resources. This is particularly true for camera-LiDAR based multimodal fusion, which may require three separate deep-learning networks and/or processing pipelines that are designated for the visual data, LiDAR data, and for some form of a fusion framework. In this paper, we propose Fast Camera-LiDAR Object Candidates (Fast-CLOCs) fusion network that can run high-accuracy fusion-based 3D object detection in near real-time. Fast-CLOCs operates on the output candidates before Non-Maximum Suppression (NMS) of any 3D detector, and adds a lightweight 3D detector-cued 2D image detector (3D-Q-2D) to extract visual features from the image domain to improve 3D detections significantly. The 3D detection candidates are shared with the proposed 3D-Q-2D image detector as proposals to reduce the network complexity drastically. The superior experimental results of our Fast-CLOCs on the challenging KITTI and nuScenes dataset s illustrate that our Fast-CLOCs outperforms state-of-the-art fusion-based 3D object detection approaches.

****

Typenet: Towards Camera Enabled Touch Typing on Flat Surfaces Through Self-Refinement

Ben Maman, Amit Bermano; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1140-1149

Text entry for mobile devices nowadays is an equally crucial and time-consuming task, with no practical solution available for natural typing speeds without extra hardware. In this paper, we introduce a real-time method that is a significant step towards enabling touch typing on arbitrary flat surfaces (e.g., tables). The method employs only a simple video camera, placed in front of the user on the flat surface --- at an angle practical for mobile usage. To achieve this, we adopt a classification framework, based on the observation that, in touch typing, similar hand configurations imply the same typed character across users. Importantly, this approach allows the convenience of un-calibrated typing, where the hand positions, with respect to the camera and each other, are not dictated. To improve accuracy, we propose a Language Processing scheme, which corrects the typed text and is specifically designed for real-time performance and integration with the vision-based signal. To enable feasible data collection and training, we propose a self-refinement approach that allows training on unlabeled flat-surface-typing footage; A network trained on (labeled) keyboard footage labels flat-surface videos using dynamic time warping, and is trained on them, in an Expectation Maximization (EM) manner. Using these techniques, we introduce the TypingHands26 Dataset, comprising videos of 26 different users typing on a keyboard, and 10 users typing on a flat surface, labeled at the frame level. We validate our approach and present a single camera-based system with character-level accuracy of 93.5% on average for known users, and 85.7% for unknown ones, outperforming pose-estimation-based methods by a large margin, despite performing at natural typing speeds of up to 80 Words Per Minute. Our method is the first to rely on a simple camera alone, and runs in interactive speeds, while still maintaining accuracy comparable to systems employing non-commodity equipment.

****

Danish Fungi 2020 - Not Just Another Image Recognition Dataset

Lukáš Picek, Milan Šulc, Ji■í Matas, Thomas S. Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, Tobias Frøslev; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1525-1535

We introduce a novel fine-grained dataset and benchmark, the Danish Fungi 2020 (DF20). The dataset, constructed from observations submitted to the Atlas of Danish Fungi, is unique in its taxonomy-accurate class labels, small number of errors, highly unbalanced long-tailed class distribution, rich observation metadata, and well-defined class hierarchy. DF20 has zero overlap with ImageNet, allowing unbiased comparison of models fine-tuned from publicly available ImageNet checkpoints. The proposed evaluation protocol enables testing the ability to improve classification using metadata - e.g. precise geographic location, habitat, and substrate, facilitates classifier calibration testing, and finally allows to study the impact of the device settings on the classification performance. Experiments using Convolutional Neural Networks (CNN) and the recent Vision Transformers (ViT) show that DF20 presents a challenging task. Interestingly, ViT achieves results superior to CNN baselines with 80.45% accuracy and 0.743 macro F1 score, reducing the CNN error by 9% and 12% respectively. A simple procedure for including metadata into the decision process improves the classification accuracy by more than 2.95 percentage points, reducing the error rate by 15%. The source code for all methods and experiments is available at https://sites.google.com/view/danish-fungi-dataset.

****

Single Image Deraining Network With Rain Embedding Consistency and Layered LSTM

Yizhou Li, Yusuke Monno, Masatoshi Okutomi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 4060-4069

Single image deraining is typically addressed as residual learning to predict th e rain layer from an input rainy image. For this purpose, an encoder-decoder network draws wide attention, where the encoder is required to encode a high-quality rain embedding which determines the performance of the subsequent decoding stage to reconstruct the rain layer. However, most of existing studies ignore the significance of rain embedding quality, thus leading to limited performance with over/under-deraining. In this paper, with our observation of the high rain layer reconstruction performance by an rain-to-rain autoencoder, we introduce the idea of "Rain Embedding Consistency" by regarding the encoded embedding by the autoencoder as an ideal rain embedding and aim at enhancing the deraining performance by improving the consistency between the ideal rain embedding and the rain embedding derived by the encoder of the deraining network. To achieve this, a Rain Embedding Loss is applied to directly supervise the encoding process, with a Rectified Local Contrast Normalization (RLCN) as the guide that effectively extracts the candidate rain pixels. We also propose Layered LSTM for recurrent deraining and fine-grained encoder feature refinement considering different scales. Qualitative and quantitative experiments demonstrate that our proposed method outperforms previous state-of-the-art methods particularly on a real-world dataset. Our source code is available at http://www.ok.sc.e.titech.ac.jp/res/SIR/.

****

Nonnegative Low-Rank Tensor Completion via Dual Formulation With Applications to Image and Video Completion

Tanmay Kumar Sinha, Jayadev Naram, Pawan Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3732-3740

Recent approaches to the tensor completion problem have often overlooked the nonnegative structure of the data. We consider the problem of learning a nonnegative low-rank tensor, and using duality theory, we propose a novel factorization of such tensors. The factorization decouples the nonnegative constraints from the low-rank constraints. The resulting problem is an optimization problem on manifolds, and we propose a variant of Riemannian conjugate gradients to solve it. We test the proposed algorithm across various tasks such as colour image inpainting, video completion, and hyperspectral image completion. Experimental results show that the proposed method outperforms many state-of-the-art tensor completion algorithms.

****

On Black-Box Explanation for Face Verification

Domingo Mery, Bernardita Morris; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3418-3427

Given a facial matcher, in explainable face verification, the task is to answer: how relevant are the parts of a probe image to establish the matching with an enrolled image. In many cases, however, the trained models cannot be manipulated and must be treated as "black-boxes". In this paper, we present six different saliency maps that can be used to explain any face verification algorithm with no manipulation inside of the face recognition model. The key idea of the methods is based on how the matching score of the two face images changes when the probe is perturbed. The proposed methods remove and aggregate different parts of the face, and measure contributions of these parts individually and in-collaboration as well. We test and compare our proposed methods in three different scenarios: synthetic images with different qualities and occlusions, real face images with different facial expressions, poses, and occlusions and faces from different demographic groups. In our experiments, five different face verification algorithms are used: ArcFace, Dlib, FaceNet (trained on VGGface2 and Casia-WebFace), and LBP. We conclude that one of the proposed methods achieves saliency maps that are stable and interpretable to humans. In addition, our method, in combination with a new visualization of saliency maps based on contours, shows promising results in comparison with other state-of-the-art art methods. This paper presents good insights into any face verification algorithm, in which it can be clearly appreciated which are the most relevant face areas that an algorithm takes into account to carry out the recognition process.

****

Challenges in Procedural Multimodal Machine Comprehension: A Novel Way To Benchm ark

Pritish Sahu, Karan Sikka, Ajay Divakaran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3654-3663

We focus on Multimodal Machine Reading Comprehension (M3C) where a model is expected to answer questions based on given passage (or context), and the context and the questions can be in different modalities. Previous works such as RecipeQA have proposed datasets and cloze-style tasks for evaluation. However, we identify three critical biases stemming from the question-answer generation process and memorization capabilities of large deep models. These biases makes it easier for a model to overfit by relying on spurious correlations or naive data patterns. We propose a systematic framework to address these biases through three Control-Knobs that enable us to generate a test bed of datasets of progressive difficulty levels. We believe that our benchmark (referred to as Meta- RecipeQA) will provide, for the first time, a fine grained estimate of a model's generalization capabilities. We also propose a generalM3C model that is used to realize several prior SOTA models and motivate a novel hierarchical transformer based reasoning network (HTRN). We perform a detailed evaluation of these models with different language and visual features on our benchmark. We observe a consistent improvement with HTRN over SOTA (18% in Visual Cloze task and   13% in average over all the tasks). We also observe a drop in performance across all the models when testing on RecipeQA and proposed Meta-RecipeQA (e.g. 83.6% versus 67.1% for HTRN), which shows that the proposed dataset is relatively less biased. We conclude by highlighting the impact of the control knobs with some quantitative results.

****

GANs Spatial Control via Inference-Time Adaptive Normalization

Karin Jakoel, Liron Efraim, Tamar Rott Shaham; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2160-2169

We introduce a new approach for spatial control over the generation process of Generative Adversarial Networks (GANs). Our approach includes modifying the normalization scheme of a pre-trained GAN at test time, so as to act differently at different image regions, according to guidance from the user. This enables to achieve different generation effects at different locations across the image. In contrast to previous works that require either fine-tuning the model's parameters or training an additional network, our approach uses the pre-trained GAN as is, without any further modifications or training phase. Our method is thus completely generic and can be easily incorporated into common GAN models. We prove our technique to be useful for solving a line of image manipulation tasks, allowing different generation effects across the image, while preserving the GAN's high visual quality.

****

SSCAP: Self-Supervised Co-Occurrence Action Parsing for Unsupervised Temporal Action Segmentation

Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, Charless Fowlkes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1819-1828

Temporal action segmentation is a task to classify each frame in the video with an action label. However, it is quite expensive to annotate every frame in a large corpus of videos to construct a comprehensive supervised training dataset. Thus in this work we propose an unsupervised method, namely SSCAP, that operates on a corpus of unlabeled videos and predicts a likely set of temporal segments across the videos. SSCAP leverages Self-Supervised learning to extract distinguishable features and then applies a novel Co-occurrence Action Parsing algorithm to not only capture the correlation among sub-actions underlying the structure of activities, but also estimate the temporal path of the sub-actions in an accurate and general way. We evaluate on both classic datasets (Breakfast, 50Salads) and the emerging fine-grained action dataset (FineGym) with more complex activity structures and similar sub-actions. Results show that SSCAP achieves state-of-the-art performance on all datasets and can even outperform some weakly-supervised approaches, demonstrating its effectiveness and generalizability.

****

Equine Pain Behavior Classification via Self-Supervised Disentangled Pose Representation

Maheen Rashid, Sofia Broomé, Katrina Ask, Elin Hernlund, Pia Haubro Andersen, Hedvig Kjellström, Yong Jae Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1646-1656

Timely detection of horse pain is important for equine welfare. Horses express pain through their facial and body behavior, but may hide signs of pain from unfamiliar human observers. In addition, collecting visual data with detailed annotation of horse behavior and pain state is both cumbersome and not scalable. Consequently, a pragmatic equine pain classification system would use video of the unobserved horse and weak labels. This paper proposes such a method for equine pain classification by using multi-view surveillance video footage of unobserved horses with induced orthopaedic pain, with temporally sparse video level pain labels. To ensure that pain is learned from horse body language alone, we first train a self-supervised generative model to disentangle horse pose from its appearance and background before using the disentangled horse pose latent representation for pain classification. To make best use of the pain labels, we develop a novel loss that formulates pain classification as a multi-instance learning problem. Our method achieves pain classification accuracy better than human expert performance with 60% accuracy. The learned latent horse pose representation is shown to be viewpoint covariant, and disentangled from horse appearance. Qualitative analysis of pain classified segments shows correspondence between the pain symptoms identified by our model, and equine pain scales used in veterinary practice.

****

PICA: Point-Wise Instance and Centroid Alignment Based Few-Shot Domain Adaptive Object Detection With Loose Annotations

Chaoliang Zhong, Jie Wang, Cheng Feng, Ying Zhang, Jun Sun, Yasuto Yokota; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2329-2338

In this work, we focus on supervised domain adaptation for object detection in few-shot loose annotation setting, where the source images are sufficient and fully labeled but the target images are few-shot and loosely annotated. As annotated objects exist in the target domain, instance level alignment can be utilized to improve the performance. Traditional methods conduct the instance level alignment by semantically aligning the distributions of paired object features with domain adversarial training. Although it is demonstrated that point-wise surrogates of distribution alignment provide a more effective solution in few-shot classification tasks across domains, this point-wise alignment approach has not yet been extended to object detection. In this work, we propose a method that extends the point-wise alignment from classification to object detection. Moreover, in the few-shot loose annotation setting, the background ROIs of target domain suffer from severe label noise problem, which may make the point-wise alignment fail. To this end, we exploit moving average centroids to mitigate the label noise problem of background ROIs. Meanwhile, we exploit point-wise alignment over instances and centroids to tackle the problem of scarcity of labeled target instances. Hence this method is not only robust against label noises of background ROIs but also robust against the scarcity of labeled target objects. Experimental results show that the proposed instance level alignment method brings significant improvement compared with the baseline and is superior to state-of-the-art methods.

****

Latent Reweighting, an Almost Free Improvement for GANs

Thibaut Issenhuth, Ugo Tanielian, David Picard, Jérémie Mary; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1668-1677

Standard formulations of GANs, where a continuous function deforms a connected latent space, have been shown to be misspecified when fitting different classes of images. In particular, the generator will necessarily sample some low-quality images in between the classes. Rather than modifying the architecture, a line of works aims at improving the sampling quality from pre-trained generators at the expense of increased computational cost. Building on this, we introduce an addi tional network to predict latent importance weights and two associated sampling methods to avoid the poorest samples. This idea has several advantages: 1) it provides a way to inject disconnectedness into any GAN architecture, 2) since the rejection happens in the latent space, it avoids going through both the generator and the discriminator, saving computation time, 3) this importance weights formulation provides a principled way to reduce the Wasserstein's distance to the target distribution. We demonstrate the effectiveness of our method on several datasets, both synthetic and high-dimensional.

****

3DFaceFill: An Analysis-by-Synthesis Approach To Face Completion
Rahul Dey, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1586-1595

Existing face completion solutions are primarily driven by end-to-end models that directly generate 2D completions of 2D masked faces. By having to implicitly account for geometric and photometric variations in facial shape and appearance, such approaches result in unrealistic completions, especially under large variations in pose, shape, illumination and mask sizes. To alleviate these limitations, we introduce 3DFaceFill, an analysis-by-synthesis approach for face completion that explicitly considers the image formation process. It comprises three components, (1) an encoder that disentangles the face into its constituent 3D mesh, 3D pose, illumination and albedo factors, (2) an autoencoder that inpaints the UV representation of facial albedo, and (3) a renderer that resynthesizes the completed face. By operating on the UV representation, 3DFaceFill affords the power of correspondence and allows us to naturally enforce geometrical priors (e.g. facial symmetry) more effectively. Quantitatively, 3DFaceFill improves the state-of-the-art by up to 4dB higher PSNR and 25% better LPIPS for large masks. And, qualitatively, it leads to demonstrably more photorealistic face completions over a range of masks and occlusions while preserving consistency in global and component-wise shape, pose, illumination and eye-gaze.

****

MisConv: Convolutional Neural Networks for Missing Data
Marcin Przewi■■likowski, Marek ■mieja, ■ukasz Struski, Jacek Tabor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2060-2069

Processing of missing data by modern neural networks, such as CNNs, remains a fundamental, yet unsolved challenge, which naturally arises in many practical applications, like image inpainting or autonomous vehicles and robots. While imputation-based techniques are still one of the most popular solutions, they frequently introduce unreliable information to the data and do not take into account the uncertainty of estimation, which may be destructive for a machine learning model. In this paper, we present MisConv, a general mechanism, for adapting various CNN architectures to process incomplete images. By modeling the distribution of missing values by the Mixture of Factor Analyzers, we cover the spectrum of possible replacements and find an analytical formula for the expected value of convolution operator applied to the incomplete image. The whole framework is realized by matrix operations, which makes MisConv extremely efficient in practice. Experiments performed on various image processing tasks demonstrate that MisConv achieves superior or comparable performance to the state-of-the-art methods.

****

Surrogate Model-Based Explainability Methods for Point Cloud NNs
Hanxiao Tan, Helena Kotthaus; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2239-2248

In the field of autonomous driving and robotics, point clouds are showing their excellent real-time performance as raw data from most of the mainstream 3D sensors. Therefore, point cloud neural networks have become a popular research direction in recent years. So far, however, there has been little discussion about the explainability of deep neural networks for point clouds. In this paper, we propose a point cloud-applicable explainability approaches based on local surrogate model-based methods to show which components make the main contribution to the classification. Moreover, we propose quantitative fidelity validations for genera ted explanations that enhance the persuasive power of explainability and compare the plausibility of different existing point cloud-applicable explainability methods. Our new explainability approach provides a fairly accurate, more intuitive and widely applicable explanation for point cloud classification tasks. Our code is available at https://github.com/Explain3D/LIME-3D

****

3D Modeling Beneath Ground: Plant Root Detection and Reconstruction Based on Ground-Penetrating Radar

Yawen Lu, Guoyu Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 68-77

3D object reconstruction based on deep neural networks has been gaining attention in recent years. However, recovering 3D shapes of hidden and buried objects remains to be a challenge. Ground Penetrating Radar (GPR) is among the most powerful and widely used instruments for detecting and locating underground objects such as plant roots and pipes, with affordable prices and continually evolving technology. This paper first proposes a deep convolution neural network-based anchor-free GPR curve signal detection network utilizing B-scans from a GPR sensor. The detection results can help obtain precisely fitted parabola curves. Furthermore, a graph neural network-based root shape reconstruction network is designated in order to progressively recover major taproot and then fine root branches' geometry. Our results on the gprMax simulated root data as well as the real-world GPR data collected from apple orchards demonstrate the potential of using the proposed framework as a new approach for fine-grained underground object shape reconstruction in a non-destructive way.

****

Identifying Wrongly Predicted Samples: A Method for Active Learning

Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2290-2298

While unlabelled data can be largely available and even abundant, the annotation process can be quite expensive and limiting. Under the assumption that some samples are more important for a given task than others, active learning targets the problem of identifying the most informative samples that one should acquire annotations for. In this work we propose a simple sample selection criterion that moves beyond the conventional reliance on model uncertainty as proxy to leverage new labels. By first accepting the model prediction and then judging its effect on the generalization error, we can better identify wrongly predicted samples. We also present a very efficient approximation to our criterion, providing a similarity-based interpretation. In addition to evaluating our method on the standard benchmarks of active learning, we consider the challenging yet realistic imbalanced data scenario. We show state-of-the-art results, especially on the imbalanced setting, and achieve better rates at identifying wrongly predicted samples than existing active learning methods. Our method is simple, model agnostic and relies on the current model status without the need for re-training from scratch.

****

Learning Color Representations for Low-Light Image Enhancement

Bomi Kim, Sunhyeok Lee, Nahyun Kim, Donggon Jang, Dae-Shik Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1455-1463

Color conveys important information about the visible world. However, under low-light conditions, both pixel intensity, as well as true color distribution, can be significantly shifted. Moreover, most of such distortions are non-recoverable due to inverse problems. In the present study, we utilized recent advancements in learning-based methods for low-light image enhancement. However, while most "deep learning" methods aim to restore high-level and object-oriented visual information, we hypothesized that learning-based methods can also be used for restoring color-based information. To address this question, we propose a novel color representation learning method for low-light image enhancement. More specifically, we used a channel-aware residual network and a differentiable intensity histo gram to capture color features. Experimental results using synthetic and natural datasets suggest that the proposed learning scheme achieves state-of-the-art performance. We conclude from our study that inter-channel dependency and color distribution matching are crucial factors for learning color representations under low-light conditions.

****

Event-Based Kilohertz Eye Tracking Using Coded Differential Lighting

Timo Stoffregen, Hossein Daraei, Clare Robinson, Alexander Fix; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2515-2523

Pixels in an event camera operate asynchronously and independently, reporting changes in intensity as events - tuples of (x,y) position, polarity s and timestamp t at microsecond resolution. Event cameras operate at low power (5mW) and respond to changes in the scene with a latency on the order of microseconds. These properties make event cameras an exciting candidate for eye tracking sensors on mobile platforms such as AR/VR headsets, since these systems have hard real-time and power constraints. One proven method for eye tracking and gaze estimation is corneal glint detection. We exploit the fact that corneal glint tracking only requires a sparse set of pixels in the image, by making use of the natural sparsity of event cameras, which only detect changes in the scene. To enhance this effect, we design an illumination scheme, Coded Differential Lighting, which enhances specular reflections, suppresses all other events, and solves the light-to-glint correspondence. This is the first purely event-based corneal glint detection and tracking algorithm, which operates on standard hardware at kHz sampling rate.

****

DeepPatent: Large Scale Patent Drawing Recognition and Retrieval

Michal Kucer, Diane Oyen, Juan Castorena, Jian Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2309-2318

We tackle the problem of analyzing and retrieving technical drawings. First, we introduce DeepPatent, a new large-scale dataset for recognition and retrieval of design patent drawings. The dataset provides more than 350,000 design patent drawings for the purpose of image retrieval. Unlike existing datasets, DeepPatent provides fine-grained image retrieval associations within the collection of drawings and does not rely on cross-domain associations for supervision. We develop a baseline deep learning models, named PatentNet, based on best practices for training retrieval models for static images. We demonstrate the superior performance of PatentNet when trained on our fine-grained associations of DeepPatent against other deep learning approaches and classic computer vision descriptors, such as histogram of oriented gradients (HOG), on DeepPatent. With the introduction of this new dataset, and benchmark algorithms, we demonstrate that the analysis and retrieval of line drawings remains an open challenge in computer vision; and that patent drawing retrieval provides a concrete testbench to spur research.

****

Creating and Reenacting Controllable 3D Humans With Differentiable Rendering

Thiago L. Gomes, Thiago M. Coutinho, Rafael Azevedo, Renato Martins, Erickson R. Nascimento; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1687-1696

This paper proposes a new end-to-end neural rendering architecture to transfer appearance and reenact human actors. Our method leverages a carefully designed graph convolutional network (GCN) to model the human body manifold structure, jointly with differentiable rendering, to synthesize new videos of people in different contexts from where they were initially recorded. Unlike recent appearance transferring methods, our approach can reconstruct a fully controllable 3D texture-mapped model of a person, while taking into account the manifold structure from body shape and texture appearance in the view synthesis. Specifically, our approach models mesh deformations with a three-stage GCN trained in a self-supervised manner on rendered silhouettes of the human body. It also infers texture appearance with a convolutional network in the texture domain, which is trained in an adversarial regime to reconstruct human texture from rendered images of actors in different poses. Experiments on different videos show that our method successfully infers specific body deformations and avoid creating texture artifacts while achieving the best values for appearance in terms of Structural Similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Mean Squared Error (MSE), and Frechet Video Distance (FVD). By taking advantages of both differentiable rendering and the 3D parametric model, our method is fully controllable, which allows controlling the human synthesis from both pose and rendering parameters. The source code is available at https://www.verlab.dcc.ufmg.br/retargeting-motion/wacv2022.

****

Improving Object Detection by Label Assignment Distillation

Chuong H. Nguyen, Thuy C. Nguyen, Tuan N. Tang, Nam L.H. Phan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1005-1014

Label assignment in object detection aims to assign targets, foreground or background, to sampled regions in an image. Unlike labeling for image classification, this problem is not well defined due to the object's bounding box. In this paper, we investigate the problem from a perspective of distillation, hence we call Label Assignment Distillation (LAD). Our initial motivation is very simple, we use a teacher network to generate labels for the student. This can be achieved in two ways: either using the teacher's prediction as the direct targets (soft label), or through the hard labels dynamically assigned by the teacher (LAD). Our experiments reveal that: (i) LAD is more effective than soft-label, but they are complementary. (ii) Using LAD, a smaller teacher can also improve a larger student significantly, while soft-label can't. We then introduce Co-learning LAD, in which two networks simultaneously learn from scratch and the role of teacher and student are dynamically interchanged. Using PAA-ResNet50 as a teacher, our LAD techniques can improve detectors PAA-ResNet101 and PAA-ResNeXt101 to 46 AP and 47.5 AP on the COCO test-dev set. With a stronger teacher PAA-SwinB, we improve the students PAA-ResNet50 to 43.7 AP by only 1x schedule training and standard setting, and PAA-ResNet101 to 47.9 AP, significantly surpassing the current methods. Our source code is released at https://git.io/JrDZo.

****

MAPS: Multimodal Attention for Product Similarity

Nilotpal Das, Aniket Joshi, Promod Yenigalla, Gourav Agrwal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3338-3346

Learning to identify similar products in the e-commerce domain has widespread applications such as ensuring consistent grouping of the products in the catalog, avoiding duplicates in the search results, etc. Here, we address the problem of learning product similarity for highly challenging real-world data from the Amazon catalog. We define it as a metric learning problem, where similar products are projected close to each other and dissimilar ones are projected further apart. To this end, we propose a scalable end-to-end multimodal framework for product representation learning in a weakly supervised setting using raw data from the catalog. This includes product images as well as textual attributes like product title and category information. The model uses the image as the primary source of information, while the title helps the model focus on relevant regions in the image by ignoring the background clutter. To validate our approach, we created multimodal datasets covering three broad product categories, where we achieve up to 10% improvement in precision compared to state-of-the-art multimodal benchmark. Along with this, we also incorporate several effective heuristics for training data generation, which further complements the overall training. Additionally, we demonstrate that incorporating the product title makes the model scale effectively across multiple product categories.

****

Physical Adversarial Attacks on an Aerial Imagery Object Detector

Andrew Du, Bo Chen, Tat-Jun Chin, Yee Wei Law, Michele Sasdelli, Ramesh Rajasegaran, Dillon Campbell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1796-1806

Deep neural networks (DNNs) have become essential for processing the vast amounts of aerial imagery collected using earth-observing satellite platforms. However, DNNs are vulnerable towards adversarial examples, and it is expected that this weakness also plagues DNNs for aerial imagery. In this work, we demonstrate one of the first efforts on physical adversarial attacks on aerial imagery, whereby adversarial patches were optimised, fabricated and installed on or near target objects (cars) to significantly reduce the efficacy of an object detector applied on overhead images. Physical adversarial attacks on aerial images, particularly those captured from satellite platforms, are challenged by atmospheric factors (lighting, weather, seasons) and the distance between the observer and target. To investigate the effects of these challenges, we devised novel experiments and metrics to evaluate the efficacy of physical adversarial attacks against object detectors in aerial scenes. Our results indicate the palpable threat posed by physical adversarial attacks towards DNNs for processing satellite imagery.

****

Weakly Supervised Learning for Joint Image Denoising and Protein Localization in Cryo-Electron Microscopy

Qinwen Huang, Ye Zhou, Hsuan-Fu Liu, Alberto Bartesaghi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3246-3255

Deep learning-based object detection methods have shown promising results in various fields ranging from autonomous driving to video surveillance where input images have relatively high signal-to-noise ratios (SNR). On low SNR images such as biological electron microscopy (EM) data, however, the performance of these algorithms is significantly lower. Moreover, biological data typically lacks standardized annotations further complicating the training of detection algorithms. Accurate identification of proteins from EM images is a critical task, as the detected positions serve as inputs for the downstream 3D structure determination process. To overcome the low SNR and lack of image annotations, we propose a joint weakly-supervised learning framework that performs image denoising while detecting objects of interest. By leveraging per-pixel soft segmentation and consistency regularization, our framework denoises images without the need of clean images and is able to detect particles of interest even when less than 0.5% of the data are labeled. We validate our approach on real single-particle cryo-EM and cryo-electron tomography (ET) images which are known to suffer from extremely low SNR, and show that our strategy outperforms existing state-of-the-art (SofA) methods used in the cryo-EM field by a significant margin. We also evaluate the performance of our algorithm under decreasing SNR conditions and show that our method is more robust to noise than competing methods.

****

Self-Supervised Domain Adaptation for Visual Navigation With Global Map Consistency

Eun Sun Lee, Junho Kim, Young Min Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1707-1716

We propose a light-weight, self-supervised adaptation for a visual navigation agent to generalize to unseen environment. Given an embodied agent trained in a noiseless environment, our objective is to transfer the agent to a noisy environment where actuation and odometry sensor noise is present. Our method encourages the agent to maximize the consistency between the global maps generated at different time steps in a round-trip trajectory. The proposed task is completely self-supervised, not requiring any supervision from ground-truth pose data or explicit noise model. In addition, optimization of the task objective is extremely light-weight, as training terminates within a few minutes on a commodity GPU. Our experiments show that the proposed task helps the agent to successfully transfer to new, noisy environments. The transferred agent exhibits improved localization and mapping accuracy, further leading to enhanced performance in downstream visual navigation tasks. Moreover, we demonstrate test-time adaptation with our self-supervised task to show its potential applicability in real-world deployment.

****

StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation

Umut Kocasari, Alara Dirik, Mert Tiftikci, Pinar Yanardag; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 895-904

Discovering meaningful directions in the latent space of GANs to manipulate semantic attributes typically requires large amounts of labeled data. Recent work aims to overcome this limitation by leveraging the power of Contrastive Language-Image Pre-training (CLIP), a joint text-image model. While promising, these methods require several hours of preprocessing or training to achieve the desired manipulations. In this paper, we present StyleMC, a fast and efficient method for text-driven image generation and manipulation. StyleMC uses a CLIP-based loss and an identity loss to manipulate images via a single text prompt without significantly affecting other attributes. Unlike prior work, StyleMC requires only a few seconds of training per text prompt to find stable global directions, does not require prompt engineering and can be used with any pre-trained StyleGAN2 model. We demonstrate the effectiveness of our method and compare it to state-of-the-art methods.

****

One-Shot Compositional Data Generation for Low Resource Handwritten Text Recognition

Mohamed Ali Souibgui, Ali Furkan Biten, Sounak Dey, Alicia Fornés, Yousri Kessentini, Lluís Gómez, Dimosthenis Karatzas, Josep Lladós; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 935-943

Low resource Handwritten Text Recognition (HTR) is a hard problem due to the scarce annotated data and the very limited linguistic information (dictionaries and language models). For example, in the case of historical ciphered manuscripts, which are usually written with invented alphabets to hide the message contents. Thus, in this paper we address this problem through a data generation technique based on Bayesian Program Learning (BPL). Contrary to traditional generation approaches, which require a huge amount of annotated images, our method is able to generate human-like handwriting using only one sample of each symbol in the alphabet. After generating symbols, we create synthetic lines to train state-of-the-art HTR architectures in a segmentation free fashion. Quantitative and qualitative analyses were carried out and confirm the effectiveness of the proposed method.

****

PhotoWCT2: Compact Autoencoder for Photorealistic Style Transfer Resulting From Blockwise Training and Skip Connections of High-Frequency Residuals

Tai-Yin Chiu, Danna Gurari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2868-2877

Photorealistic style transfer is an image editing task with the goal to modify an image to match the style of another image while ensuring the result looks like a real photograph. A limitation of existing models is that they have many parameters, which in turn prevents their use for larger image resolutions and leads to slower run-times. We introduce two mechanisms that enable our design of a more compact model that we call PhotoWCT2, which preserves state-of-art stylization strength and photorealism. First, we introduce blockwise training to perform coarse-to-fine feature transformations that enable state-of-art stylization strength in a single autoencoder in place of the inefficient cascade of four autoencoders used in PhotoWCT. Second, we introduce skip connections of high-frequency residuals in order to preserve image quality when applying the sequential coarse-to-fine feature transformations. Our PhotoWCT2 model requires fewer parameters (e.g., 30.3% fewer) while supporting higher resolution images (e.g., 4K) and achieving faster stylization than existing models.

****

Enhancing Few-Shot Image Classification With Unlabelled Examples

Peyman Bateni, Jarred Barber, Jan-Willem van de Meent, Frank Wood; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2796-2805

We develop a transductive meta-learning method that uses unlabelled instances to

improve few-shot image classification performance. Our approach combines a regularized Mahalanobis-distance-based soft k-means clustering procedure with a modified state of the art neural adaptive feature extractor to achieve improved test-time classification accuracy using unlabelled data. We evaluate our method on transductive few-shot learning tasks, in which the goal is to jointly predict labels for query (test) examples given a set of support (training) examples. We achieve state of the art performance on the Meta-Dataset, mini-ImageNet and tiered-ImageNet benchmarks. All trained models and code have been made publicly available at github.com/plai-group/simple-cnaps.
****
Masking Modalities for Cross-Modal Video Retrieval
Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1766-1775
Pre-training on large scale unlabelled datasets has shown impressive performance improvements in the fields of computer vision and natural language processing. Given the advent of large-scale instructional video datasets, a common strategy for pre-training video encoders is to use the accompanying speech as weak supervision. However, as speech is used to supervise the pre-training, it is never seen by the video encoder, which does not learn to process that modality. We address this drawback of current pre-training methods, which fail to exploit the rich cues in spoken language. Our proposal is to pre-train a video encoder using all the available video modalities as supervision, namely, appearance, sound, and transcribed speech. We mask an entire modality in the input and predict it using the other two modalities. This encourages each modality to collaborate with the others, and our video encoder learns to process appearance and audio as well as speech. We show the superior performance of our modality masking pre-training approach for video retrieval on the How2R, YouCook2 and Condensed Movies datasets.
****
Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization
Royston Rodrigues, Masahiro Tani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3871-3879
When we humans recognize places from images, we not only infer about the objects that are available but even think about landmarks that might be surrounding it. Current place recognition approaches lack the ability to go beyond objects that are available in the image and hence miss out on understanding the scene completely. In this paper, we take a step towards holistic scene understanding. We address the problem of image geo-localization by retrieving corresponding aerial views from a large database of geotagged aerial imagery. One of the main challenges in tackling this problem is the limited Field of View (FoV) nature of query images which needs to be matched to aerial views which contain 360degFoV details. State-of-the-art method DSM-Net [17] tackles this challenge by matching aerial images locally within fixed FoV sectors. We show that local matching limits complete scene understanding and is inadequate when partial buildings are visible in query images or when local sectors of aerial images are covered by dense trees. Our approach considers both local and global properties of aerial images and hence is robust to such conditions. Experiments on standard benchmarks demonstrates that the proposed approach improves top-1% image recall rate on the CVACT [9] data-set from 57.08% to 77.19% and from 61.20% to 75.21% on the CVUSA [25] data-set for 70degFoV. We also achieve state-of-the art results for 90degFoV on both CVACT [9] and CVUSA [25] data-sets demonstrating the effectiveness of our proposed method.
****
StickyLocalization: Robust End-to-End Relocalization on Point Clouds Using Graph Neural Networks
Kai Fischer, Martin Simon, Stefan Milz, Patrick Mäder; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2962-2971
Relocalization inside pre-built maps provides a big benefit in the course of tod

ay's autonomous driving tasks where the map can be considered as an additional sensor for refining the estimated current pose of the vehicle. Due to potentially large drifts in the initial pose guess as well as maps containing unfiltered dynamic and temporal static objects (e.g. parking cars), traditional methods like ICP tend to fail and show high computation times. We propose a novel and fast relocalization method for accurate pose estimation inside a pre-built map based on 3D point clouds. The method is robust against inaccurate initialization caused by low performance GPS systems and tolerates the presence of unfiltered objects by specifically learning to extract significant features from current scans and adjacent map sections. More specifically, we introduce a novel distance-based matching loss enabling us to simultaneously extract important information from raw point clouds and aggregating inner- and inter-cloud context by utilizing self-and cross-attention inside a Graph Neural Network. We evaluate StickyLocalization's (SL) performance through an extensive series of experiments using two benchmark datasets in terms of Relocalization on NuScenes and Loop Closing using KITTI's Odometry dataset. We found that SL outperforms state-of-the art point cloud registration and relocalization methods in terms of transformation errors and run time.

****

Siamese Transformer Pyramid Networks for Real-Time UAV Tracking

Daitao Xing, Nikolaos Evangeliou, Athanasios Tsoukalas, Anthony Tzes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2139-2148

Recent object tracking methods depend upon deep networks or convoluted architectures. Most of those trackers can hardly meet real-time processing requirements on mobile platforms with limited computing resources. In this work, we introduce the Siamese Transformer Pyramid Network (SiamTPN), which inherits the advantages from both CNN and Transformer architectures. Specifically, we exploit the inherent feature pyramid of a lightweight network (ShuffleNetV2) and reinforce it with a Transformer to construct a robust target-specific appearance model. A centralized architecture with lateral cross attention is developed for building augmented high-level feature maps. To avoid the computation and memory intensity while fusing pyramid representations with the Transformer, we further introduce the pooling attention module, which significantly reduces memory and time complexity while improving the robustness. Comprehensive experiments on both aerial and prevalent tracking benchmarks achieve competitive results while operating at high speed, demonstrating the effectiveness of SiamTPN. Moreover, our fastest variant tracker operates over 30 Hz on a single CPU-core and obtaining an AUC score of 58.1% on the LaSOT dataset.

****

Automated Defect Inspection in Reverse Engineering of Integrated Circuits

Ann-Christin Bette, Patrick Brus, Gabor Balazs, Matthias Ludwig, Alois Knoll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1596-1605

In the semiconductor industry, reverse engineering is used to extract information from microchips. Circuit extraction is becoming increasingly difficult due to the continuous technology shrinking. A high quality reverse engineering process is challenged by various defects coming from chip preparation and imaging errors. Currently, no automated, technology-agnostic defect inspection framework is available. To meet the requirements of the mostly manual reverse engineering process, the proposed automated framework needs to handle highly imbalanced data, as well as unknown and multiple defect classes. We propose a network architecture that is composed of a shared Xception-based feature extractor and multiple, individually trainable binary classification heads: the HydREnet. We evaluated our defect classifier on three challenging industrial datasets and achieved accuracies of over 85 %, even for underrepresented classes. With this framework, the manual inspection effort can be reduced down to 5 %.

****

Improving Model Generalization by Agreement of Learned Representations From Data Augmentation

Rowel Atienza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 372-381

Data augmentation reduces the generalization error by forcing a model to learn invariant representations given different transformations of the input image. In computer vision, on top of the standard image processing functions, data augmentation techniques based on regional dropout such as CutOut, MixUp, and CutMix and policy-based selection such as AutoAugment demonstrated state-of-the-art (SOTA) results. With an increasing number of data augmentation algorithms being proposed, the focus is always on optimizing the input-output mapping while not realizing that there might be an untapped value in the transformed images with the same label. We hypothesize that by forcing the representations of two transformations to agree, we can further reduce the model generalization error. We call our proposed method Agreement Maximization or simply AgMax. With this simple constraint applied during training, empirical results show that data augmentation algorithms can further improve the classification accuracy of ResNet50 on ImageNet by up to 1.5%, WideResNet40-2 on CIFAR10 by up to 0.7%, WideResNet40-2 on CIFAR100 by up to 1.6%, and LeNet5 on Speech Commands Dataset by up to 1.4%. Experimental results further show that unlike other regularization terms such as label smoothing, AgMax can take advantage of the data augmentation to consistently improve model generalization by a significant margin. On downstream tasks such as object detection and segmentation on PascalVOC and COCO, AgMax pre-trained models outperforms other data augmentation methods by as much as 1.0mAP (box) and 0.5mAP (mask). Code is available at https://github.com/roatienza/agmax.

****

Facial Attribute Transformers for Precise and Robust Makeup Transfer

Zhaoyi Wan, Haoran Chen, Jie An, Wentao Jiang, Cong Yao, Jiebo Luo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1717-1726

In this paper, we address the problem of makeup transfer, which aims at transplanting the makeup from the reference face to the source face while preserving the identity of the source. Existing makeup transfer methods have made notable progress in generating realistic makeup faces, but do not perform well in terms of color fidelity and spatial transformation. To tackle these issues, we propose a novel Facial Attribute Transformer (FAT) and its variant Spatial FAT for high-quality makeup transfer. Drawing inspirations from the Transformer in NLP, FAT is able to model the semantic correspondences and interactions between the source face and reference face, and then precisely estimate and transfer the facial attributes. To further facilitate shape deformation and transformation of facial parts, we also integrate thin plate splines (TPS) into FAT, thus creating Spatial FAT, which is the first method that can transfer geometric attributes in addition to color and texture. Extensive qualitative and quantitative experiments demonstrate the effectiveness and superiority of our proposed FATs in the following aspects: (1) ensuring high-fidelity color transfer; (2) allowing for geometric transformation of facial parts; (3) handling facial variations (such as poses and shadows) and (4) supporting high-resolution face generation.

****

Auto White-Balance Correction for Mixed-Illuminant Scenes

Mahmoud Afifi, Marcus A. Brubaker, Michael S. Brown; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1210-1219

Auto white balance (AWB) is applied by camera hardware at capture time to remove the color cast caused by the scene illumination. The vast majority of white-balance algorithms assume a single light source illuminates the scene; however, real scenes often have mixed lighting conditions. This paper presents an effective AWB method to deal with such mixed-illuminant scenes. A unique departure from conventional AWB, our method does not require illuminant estimation, as is the case in traditional camera AWB modules. Instead, our method proposes to render the captured scene with a small set of predefined white-balance settings. Given this set of rendered images, our method learns to estimate weighting maps that are used to blend the rendered images to generate the final corrected image. Through extensive experiments, we show this proposed method produces promising results compared to other alternatives for single- and mixed-illuminant scene color correction.

****

Registration of Human Point Set Using Automatic Key Point Detection and Region-Aware Features

Amar Maharjan, Xiaohui Yuan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 741-749

Non-rigid point set registration is challenging when point sets have large deformations and different numbers of points. Examples of such point sets include human point sets representing complex human poses captured by different types of depth cameras. In this work, we present a probabilistic, non-rigid registration method to deal with these issues. Two regularization terms are used: key point correspondences and local neighborhood preservation. Our method detects key points in the point sets based on geodesic distance. Correspondences are established using a new cluster-based, region-aware feature descriptor. This feature descriptor encodes the association of a cluster to the left-right (symmetry) or upper-lower regions of the point sets. We use the Stochastic Neighbor Embedding (SNE) constraint to preserve the local neighborhood of the point set. Experimental results on challenging 3D human poses demonstrate that our method outperforms the state-of-the-art methods. Our method achieved highly competitive performance with a slight increase of error by 3.9% in comparison with the method using manually specified key point correspondences.

****

Monocular Depth Estimation With Adaptive Geometric Attention

Taher Naderi, Amir Sadovnik, Jason Hayward, Hairong Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 944-954

Single image depth estimation is an ill-posed problem. That is, it is not mathematically possible to uniquely estimate the 3rd dimension (or depth) from a single 2D image. Hence, additional constraints need to be incorporated in order to regulate the solution space. In this paper, we explore the idea of constraining the model by taking advantage of the similarity between the RGB image and the corresponding depth map at the geometric edges of the 3D scene for more accurate depth estimation. We propose a general light-weight adaptive geometric attention module that uses the cross-correlation between the encoder and the decoder as a measure of this similarity. More precisely, we use the cosine similarity between the local embedded features in the encoder and the decoder at each spatial point. The proposed module along with the encoder-decoder network is trained in an end-to-end fashion and achieves superior and competitive performance in comparison with other state-of-the-art methods. In addition, adding our module to the base encoder-decoder model adds only an additional 0.03% (or 0.0003) parameters. Therefore, this module can be added to any base encoder-decoder network without changing its structure to address any task at hand.

****

ForeSI: Success-Aware Visual Navigation Agent

Mahdi Kazemi Moghaddam, Ehsan Abbasnejad, Qi Wu, Javen Qinfeng Shi, Anton Van Den Hengel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 691-700

In this work, we present a method to improve the efficiency and robustness of the previous model-free Reinforcement Learning (RL) algorithms for the task of object-goal visual navigation. Despite achieving state-of-the-art results, one of the major drawbacks of those approaches is the lack of a forward model that informs the agent about the potential consequences of its actions, i.e., being model-free. In this work, we augment the model-free RL with such a forward model that can predict a representation of a future state, from the beginning of a navigation episode, if the episode were to be successful. Furthermore, in order for efficient training, we develop an algorithm to integrate a replay buffer into the model-free RL that alternates between training the policy and the forward model. We call our agent ForeSI; ForeSI is trained to imagine a future latent state that leads to success. By explicitly imagining such a state, during the navigation, our agent is able to take better actions leading to two main advantages: first, in the absence of an object detector, ForeSI presents a more robust policy, i.e., it leads to about 5% absolute improvement on the Success Rate (SR); second, when combined with an off-the-shelf object detector to help better distinguish the target object, our method leads to about 3% absolute improvement on the SR and about 2% absolute improvement on Success weighted by inverse Path Length (SPL), i.e., presents higher efficiency.

****

Self-Supervised Test-Time Adaptation on Video Data

Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, Andreas Dengel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3439-3448

In typical computer vision problems revolving around video data, pre-trained models are simply evaluated at test time, without adaptation. This general approach clearly cannot capture the shifts that will likely arise between the distributions from which training and test data have been sampled. Adapting a pre-trained model to a new video encountered at test time could be essential to avoid the potentially catastrophic effects of such shifts. However, given the inherent impossibility of labeling data only available at test-time, traditional fine-tuning techniques cannot be leveraged in this highly practical scenario. This paper explores whether the recent progress in test-time adaptation in the image domain and self-supervised learning can be leveraged to adapt a model to previously unseen and unlabelled videos presenting both mild (but arbitrary) and severe covariate shifts. In our experiments, we show that test-time adaptation approaches applied to self-supervised methods are always beneficial, but also that the extent of their effectiveness largely depends on the specific combination of the algorithms used for adaptation and self-supervision, and also on the type of covariate shift taking place.

****

Self-Supervised Learning of Domain Invariant Features for Depth Estimation

Hiroyasu Akada, Shariq Farooq Bhat, Ibraheem Alhashim, Peter Wonka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3377-3387

We tackle the problem of unsupervised synthetic-to-real domain adaptation for single image depth estimation. An essential building block of single image depth estimation is an encoder-decoder task network that takes RGB images as input and produces depth maps as output. In this paper, we propose a novel training strategy to force the task network to learn domain invariant representations in a self-supervised manner. Specifically, we extend self-supervised learning from traditional representation learning, which works on images from a single domain, to domain invariant representation learning, which works on images from two different domains by utilizing an image-to-image translation network. Firstly, we use an image-to-image translation network to transfer domain-specific styles between synthetic and real domains. This style transfer operation allows us to obtain similar images from the different domains. Secondly, we jointly train our task network and Siamese network with the same images from the different domains to obtain domain invariance for the task network. Finally, we fine-tune the task network using labeled synthetic and unlabeled real-world data. Our training strategy yields improved generalization capability in the real-world domain. We carry out an extensive evaluation on two popular datasets for depth estimation, KITTI and Make3D. The results demonstrate that our proposed method outperforms the state-of-the-art on all metrics, e.g. by 14.7% on Sq Rel on KITTI. The source code and model weights will be made available.

****

Transductive Weakly-Supervised Player Detection Using Soccer Broadcast Videos

Chris Andrew Gadde, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 965-974

Player detection lays the foundation for many applications in the field of sports analytics including player recognition, player tracking, and activity detectio

n. In this work we study player detection in continuous long shot broadcast videos. Broadcast match videos are easy to obtain, and detection on these videos is much more challenging. We propose a transductive approach for player detection that treats it as a domain adaptation problem. We show that instance-level domain labels are significant for sufficient adaptation in the case of soccer broadcast videos. An efficient multi-model greedy labelling scheme based on visual features is proposed to annotate domain labels on bounding box predictions made by our inductive model. We use reliable instances from the inductive model inferences to train a transductive copy of the model. We create and release a fully annotated player detection dataset comprising soccer broadcast videos from the FIFA 2018 World Cup matches to evaluate our method. Our method shows significant improvements in player detection to the baseline and existing state-of-the-art methods on our dataset. We show, on average, a 16 point improvement in mAP for soccer broadcast videos by annotating domain labels for around a 100 samples per video.

****

DG-Labeler and DGL-MOTS Dataset: Boost the Autonomous Driving Perception

Yiming Cui, Zhiwen Cao, Yixin Xie, Xingyu Jiang, Feng Tao, Yingjie Victor Chen, Lin Li, Dongfang Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 58-67

Multi-object tracking and segmentation (MOTS) is a critical task for autonomous driving applications. The existing MOTS studies face two critical challenges: 1) the published datasets inadequately capture the real-world complexity for network training to address various driving settings; 2) the working pipeline annotation tool is under-studied in the literature to improve the quality of MOTS learning examples. In this work, we introduce the DG-Labeler and DGL-MOTS dataset to facilitate the training data annotation for the MOST task and accordingly improve network training accuracy and efficiency. To the best of our knowledge, our DG-Labeler is the first tool publicly available for MOTS data annotation. DG-Labeler uses the novel Depth-Granularity Module to depict the instance spatial relations and produce fine-grained instance masks. Annotated by DG-Labeler, our DGL-MOTS dataset exceeds the prior effort (i.e., KITTI MOTS and BDD100K) in data diversity, annotation quality, and temporal representations. Results on extensive cross-dataset evaluations indicate significant performance improvements for several state-of-the-art methods trained on our DGL-MOTS dataset. We believe our DGL-MOTS Dataset and DG-Labeler hold valuable potential to boost the visual perception of future transportation. Our dataset and code are available.

****

FT-DeepNets: Fault-Tolerant Convolutional Neural Networks With Kernel-Based Duplication

Iljoo Baek, Wei Chen, Zhihao Zhu, Soheil Samii, Raj Rajkumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 975-984

Deep neural network (deepnet) applications play a crucial role in safety-critical systems such as autonomous vehicles (AVs). An AV must drive safely towards its destination, avoiding obstacles, and respond quickly when the vehicle must stop. Any transient errors in software calculations or hardware memory in these deepnet applications can potentially lead to dramatically incorrect results. Therefore, assessing and mitigating any transient errors and providing robust results are important for safety-critical systems. Previous research on this subject focused on detecting errors and then recovering from the errors by re-running the network. Other approaches were based on the extent of full network duplication such as the ensemble learning-based approach to boost system fault-tolerance by leveraging each model's advantages. However, it is hard to detect errors in a deep neural network, and the computational overhead of full redundancy can be substantial. We first study the impact of the error types and locations in deepnets. We next focus on selecting which part should be duplicated using multiple ranking methods to measure the order of importance among neurons. We find that the duplication overhead for computation and memory is a trade-off between algorithmic performance and robustness. To achieve higher robustness with less system overhead, we present two error protection mechanisms that only duplicate parts of the ne

twork from critical neurons. Finally, we substantiate the practical feasibility of our approach and evaluate the improvement in the accuracy of a deepnet in the presence of errors. We demonstrate these results using a case study with real-world applications on an Nvidia GeForce RTX 2070Ti GPU and an Nvidia Xavier embedded platform used by automotive OEMs.
****

PPCD-GAN: Progressive Pruning and Class-Aware Distillation for Large-Scale Conditional GANs Compression
Duc Minh Vo, Akihiro Sugimoto, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2436-2444
We push forward neural network compression research by exploiting a novel challenging task of large-scale conditional generative adversarial networks (GANs) compression. To this end, we propose a gradually shrinking GAN (PPCD-GAN) by introducing progressive pruning residual block (PP-Res) and class-aware distillation. The PP-Res is an extension of the conventional residual block where each convolutional layer is followed by a learnable mask layer to progressively prune network parameters as training proceeds. The class-aware distillation, on the other hand, enhances the stability of training by transferring immense knowledge from a well-trained teacher model through instructive attention maps. We train the pruning and distillation processes simultaneously on a well-known GAN architecture in an end-to-end manner. After training, all redundant parameters as well as the mask layers are discarded, yielding a lighter network while retaining the performance. We comprehensively illustrate, on ImageNet 128 x 128 dataset, PPCD-GAN reduces up to 5.2x (81%) parameters against state-of-the-arts while keeping better performance.
****

T-Net: A Resource-Constrained Tiny Convolutional Neural Network for Medical Image Segmentation
Tariq M. Khan, Antonio Robles-Kelly, Syed S. Naqvi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 644-653
In this paper, we present T-Net, a fully convolutional net-work particularly well suited for resource constrained andmobile devices, which cannot cater for the computationalresources often required by much larger networks. T-NET's design allows for dual-stream information flow both insideas well as outside of the encoder-decoder pair. Here, weuse group convolutions to increase the width of the networkand, in doing so, learn a larger number of low and inter-mediate level features. We have also employed skip connec-tions in order to keep spatial information loss to a minimum.T-Net uses a dice loss for pixel-wise classification which al-leviates the effect of class imbalance. We have performedexperiments with three different applications, retinal vesselsegmentation, skin lesion segmentation and digestive tractpolyp segmentation. In our experiments, T-Net is quite com-petitive, outperforming alternatives with two or even threeorders of magnitude more trainable parameters.
****

Short-Term Solar Irradiance Prediction From Sky Images With a Clear Sky Model
Huiyu Gao, Miaomiao Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2475-2483
Integrating the solar power into the power grid system while maintaining its stability is essential for utilising such type of clean energy widely. It renders the solar irradiance (determining the solar power) forecasting a critical task. This paper tackles the problem of solar irradiance prediction from a history of sky image sequence. Most existing machine learning methods directly regress the solar irradiance values from a historical image sequence and/or solar irradiance observations. By contrast, we propose a novel deep neural network for short-term solar irradiance forecasting by leveraging a clear sky model. In particular, we build our network structure on the vision transformer to encode the spatial as well as the temporal information in the sky video sequence. We then aim to predict the solar irradiance residual from the learned representation by explicitly using a clear sky model. We evaluated our approach extensively on the existing benchmark datasets, such as TSI880 and ASI16. Results on the nowcasting task, name

ly estimation of the solar irradiance from the observations, and the forecasting task, which is up to 4-hour ahead-of-time prediction, demonstrate the superior performance of our method compared with existing machine learning algorithms.

****

Visual Understanding of Complex Table Structures From Document Images

Sachin Raja, Ajoy Mondal, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2299-2308

Table structure recognition is necessary for a comprehensive understanding of documents. Tables in unstructured business documents are tough to parse due to the high diversity of layouts, varying alignments of contents, and the presence of empty cells. The problem is particularly difficult because of challenges in identifying individual cells using visual or linguistic contexts or both. Accurate detection of table cells (including empty cells) simplifies structure extraction and hence, it becomes the prime focus of our work. We propose a novel object-detection-based deep model that captures the inherent alignments of cells within tables and is fine-tuned for fast optimization. Despite accurate detection of cells, recognizing structures for dense tables may still be challenging because of difficulties in capturing long-range row/column dependencies in presence of multi-row/column spanning cells. Therefore, we also aim to improve structure recognition by deducing a novel rectilinear graph-based formulation. From a semantics perspective, we highlight the significance of empty cells in a table. To take these cells into account, we suggest an enhancement to a popular evaluation criterion. Finally, we introduce a modestly sized evaluation dataset with an annotation style inspired by human cognition to encourage new approaches to the problem. Our framework improves the previous state-of-the-art performance by a 2.7% average F1 score on benchmark datasets.

****

Digital and Physical-World Attacks on Remote Pulse Detection

Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin W. Bowyer, Adam Czajka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2407-2416

Remote photoplethysmography (rPPG) is a technique for estimating blood volume changes from reflected light without the need for a contact sensor. We present the first examples of presentation attacks in the digital and physical domains on rPPG from face video. Digital attacks are easily performed by adding imperceptible periodic noise to the input videos. Physical attacks are performed with illumination from visible spectrum LEDs placed in close proximity to the face, while still being difficult to perceive with the human eye. We also show that our attacks extend beyond medical applications, since the method can effectively generate a strong periodic pulse on 3D-printed face masks, which presents difficulties for pulse-based face presentation attack detection (PAD). The paper concludes with ideas for using this work to improve robustness of rPPG methods and pulse-based face PAD.

****

SporeAgent: Reinforced Scene-Level Plausibility for Object Pose Refinement

Dominik Bauer, Timothy Patten, Markus Vincze; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 654-662

Observational noise, inaccurate segmentation and ambiguity due to symmetry and occlusion lead to inaccurate object pose estimates. While depth- and RGB-based pose refinement approaches increase the accuracy of the resulting pose estimates, they are susceptible to ambiguity in the observation as they consider visual alignment. We propose to leverage the fact that we often observe static, rigid scenes. Thus, the objects therein need to be under physically plausible poses. We show that considering plausibility reduces ambiguity and, in consequence, allows poses to be more accurately predicted in cluttered environments. To this end, we extend a recent RL-based registration approach towards iterative refinement of object poses. Experiments on the LINEMOD and YCB-VIDEO datasets demonstrate the state-of-the-art performance of our depth-based refinement approach.

****

CFLOW-AD: Real-Time Unsupervised Anomaly Detection With Localization via Conditi onal Normalizing Flows

Denis Gudovskiy, Shun Ishizaka, Kazuki Kozuka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 98-107

Unsupervised anomaly detection with localization has many practical applications when labeling is infeasible and, moreover, when anomaly examples are completely missing in the train data. While recently proposed models for such data setup achieve high accuracy metrics, their complexity is a limiting factor for real-time processing. In this paper, we propose a real-time model and analytically derive its relationship to prior methods. Our CFLOW-AD model is based on a conditional normalizing flow framework adopted for anomaly detection with localization. In particular, CFLOW-AD consists of a discriminatively pretrained encoder followed by a multi-scale generative decoders where the latter explicitly estimate likelihood of the encoded features. Our approach results in a computationally and memory-efficient model: CFLOW-AD is faster and smaller by a factor of 10x than prior state-of-the-art with the same input setting. Our experiments on the MVTec dataset show that CFLOW-AD outperforms previous methods by 0.36% AUROC in detection task, by 1.12% AUROC and 2.5% AUPRO in localization task, respectively. We open-source our code with fully reproducible experiments.

****

Single Image Object Counting and Localizing Using Active-Learning

Inbar Huberman-Spiegelglas, Raanan Fattal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1310-1319

The need to count and localize repeating objects in an image arises in different scenarios, such as biological microscopy studies, production-lines inspection, and surveillance recordings analysis. The use of supervised Convolutional Neural Networks (CNNs) achieves accurate object detection when trained over large class-specific datasets. The labeling effort in this approach does not pay-off when the counting is required over few images of a unique object class. We present a new method for counting and localizing repeating objects in single-image scenarios, assuming no pre-trained classifier is available. Our method trains a CNN over a small set of labels carefully collected from the input image in few active-learning iterations. At each iteration, the latent space of the network is analyzed to extract a minimal number of user-queries that strives to both sample the in-class manifold as thoroughly as possible as well as avoid redundant labels. Compared with existing user-assisted counting methods, our active-learning iterations achieve state-of-the-art performance in terms of counting and localizing accuracy, number of user mouse clicks, and running-time. This evaluation was performed through a large user study over a wide range of image classes with diverse conditions of illumination and occlusions.

****

Reconstructing Training Data From Diverse ML Models by Ensemble Inversion

Qian Wang, Daniel Kurz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2909-2917

Model Inversion (MI), in which an adversary abuses access to a trained Machine Learning (ML) model attempting to infer sensitive information about its original training data, has attracted increasing research attention. During MI, the trained model under attack (MUA) is usually frozen and used to guide the training of a generator, such as a Generative Adversarial Network (GAN), to reconstruct the distribution prior of that model. This might cause leakage of original training samples, and if successful, the privacy of dataset subjects will be at risk if the training data contains Personally Identifiable Information (PII). Therefore, an in-depth investigation of the potentials of MI techniques is crucial for the development of corresponding defense techniques. High-quality reconstruction of training data based on a single model is challenging. However, existing MI literature does not explore targeting multiple trained models simultaneously, which may provide additional information and diverse perspectives to the adversary. In this work, we propose the ensemble inversion technique that estimates the distribution of original training data, by training a generator constrained by an ensemble (or set) of trained models with shared subjects or entities. This technique leads to noticeable improvements of the quality of the generated samples with d istinguishable features of the dataset entities compared to MI of a single model. We utilize an auxiliary dataset that's similar to the presumed training data, but we also demonstrate high quality data-free model inversion without such dataset. The impact of model diversity in the ensemble is thoroughly investigated in this work, and additional constraints are utilized to further encourage sharp predictions and high activations for the reconstructed samples, leading to more accurate reconstruction of training images.

****

Multi-Domain Incremental Learning for Semantic Segmentation

Prachi Garg, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 761-771

Recent efforts in multi-domain learning for semantic segmentation attempt to learn multiple geographical datasets in a universal, joint model. A simple fine-tuning experiment performed sequentially on three popular road scene segmentation datasets demonstrates that existing segmentation frameworks fail at incrementally learning on a series of visually disparate geographical domains. When learning a new domain, the model catastrophically forgets previously learned knowledge. In this work, we pose the problem of multi-domain incremental learning for semantic segmentation. Given a model trained on a particular geographical domain, the goal is to (i) incrementally learn a new geographical domain, (ii) while retaining performance on the old domain, (iii) given that the previous domain's dataset is not accessible. We propose a dynamic architecture that assigns universally shared, domain-invariant parameters to capture homogeneous semantic features present in all domains, while dedicated domain-specific parameters learn the statistics of each domain. Our novel optimization strategy helps achieve a good balance between retention of old knowledge (stability) and acquiring new knowledge (plasticity). We demonstrate the effectiveness of our proposed solution on domain incremental settings pertaining to real-world driving scenes from roads of Germany (Cityscapes), the United States (BDD100k), and India (IDD).

****

Multi-Task Classification of Sewer Pipe Defects and Properties Using a Cross-Task Graph Neural Network Decoder

Joakim Bruslund Haurum, Meysam Madadi, Sergio Escalera, Thomas B. Moeslund; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2806-2817

The sewerage infrastructure is one of the most important and expensive infrastructures in modern society. In order to efficiently manage the sewerage infrastructure, automated sewer inspection has to be utilized. However, while sewer defect classification has been investigated for decades, little attention has been given to classifying sewer pipe properties such as water level, pipe material, and pipe shape, which are needed to evaluate the level of sewer pipe deterioration. In this work we classify sewer pipe defects and properties concurrently and present a novel decoder-focused multi-task classification architecture Cross-Task Graph Neural Network (CT-GNN), which refines the disjointed per-task predictions using cross-task information. The CT-GNN architecture extends the traditional disjointed task-heads decoder, by utilizing a cross-task graph and unique class node embeddings. The cross-task graph can either be determined a priori based on the conditional probability between the task classes or determined dynamically using self-attention. CT-GNN can be added to any backbone and trained end-to-end at a small increase in the parameter count. We achieve state-of-the-art performance on all four classification tasks in the Sewer-ML dataset, improving defect classification and water level classification by 5.3 and 8.0 percentage points, respectively. We also outperform the single task methods as well as other multi-task classification approaches while introducing 50 times fewer parameters than previous model-focused approaches. The code and models are available at the project page http://vap.aau.dk/ctgnn.

****

Efficient Counterfactual Debiasing for Visual Question Answering

Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, Rod rigo C. Barros; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3001-3010

Despite the success of neural architectures for Visual Question Answering (VQA), several recent studies have shown that VQA models are mostly driven by superficial correlations that are learned by exploiting undesired priors within training datasets. They often lack sufficient image grounding or tend to overly-rely on textual information, failing to capture knowledge from the images. This affects their generalization to test sets with slight changes in the distribution of facts. To address such an issue, some bias mitigation methods have relied on new training procedures that are capable of synthesizing counterfactual samples by masking critical objects within the images, and words within the questions, while also changing the corresponding ground truth. We propose a novel model-agnostic counterfactual training procedure, namely Efficient Counterfactual Debiasing (ECV), in which we introduce a new negative answer-assignment mechanism that exploits the probability distribution of the answers based on their frequencies, as well as an improved counterfactual sample synthesizer. Our experiments demonstrate that ECV is a simple, computationally-efficient counterfactual sample-synthesizer training procedure that establishes itself as the new state-of-the-art for unbiased VQA.

****

Improving Single-Image Defocus Deblurring: How Dual-Pixel Images Help Through Multi-Task Learning

Abdullah Abuolaim, Mahmoud Afifi, Michael S. Brown; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1231-1239

Many camera sensors use a dual-pixel (DP) design that operates as a rudimentary light field providing two sub-aperture views of a scene in a single capture. The DP sensor was developed to improve how cameras perform autofocus. Since the DP sensor's introduction, researchers have found additional uses for the DP data, such as depth estimation, reflection removal, and defocus deblurring. We are interested in the latter task of defocus deblurring. In particular, we propose a single-image deblurring network that incorporates the two sub-aperture views into a multi-task framework. Specifically, we show that jointly learning to predict the two DP views from a single blurry input image improves the network's ability to learn to deblur the image. Our experiments show this multi-task strategy achieves +1dB PSNR improvement over state-of-the-art defocus deblurring methods. In addition, our multi-task framework allows accurate DP-view synthesis (e.g., 39dB PSNR) from the single input image. These high-quality DP views can be used for other DP-based applications, such as reflection removal. As part of this effort, we have captured a new dataset of 7,059 high-quality images to support our training for the DP-view synthesis task.

****

Hierarchical Proxy-Based Loss for Deep Metric Learning

Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Douglas Gray, Dimitris Samaras; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1859-1868

Proxy-based metric learning losses are superior to pair-based losses due to their fast convergence and low training complexity. However, existing proxy-based losses focus on learning class-discriminative features while overlooking the commonalities shared across classes which are potentially useful in describing and matching samples. Moreover, they ignore the implicit hierarchy of categories in real-world datasets, where similar subordinate classes can be grouped together. In this paper, we present a framework that leverages this implicit hierarchy by imposing a hierarchical structure on the proxies and can be used with any existing proxy-based loss. This allows our model to capture both class-discriminative features and class-shared characteristics without breaking the implicit data hierarchy. We evaluate our method on five established image retrieval datasets such as In-Shop and SOP. Results demonstrate that our hierarchical proxy-based loss framework improves the performance of existing proxy-based losses, especially on large datasets which exhibit strong hierarchical structure.

****

How Good Is Your Explanation? Algorithmic Stability Measures To Assess the Quality of Explanations for Deep Neural Networks
Thomas Fel, David Vigouroux, Rémi Cadène, Thomas Serre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 720-730
A plethora of methods have been proposed to explain how deep neural networks reach their decisions but comparatively, little effort has been made to ensure that the explanations produced by these methods are objectively relevant. While several desirable properties for trustworthy explanations have been formulated, objective measures have been harder to derive. Here, we propose two new measures to evaluate explanations borrowed from the field of algorithmic stability: mean generalizability MeGe and relative consistency ReCo. We conduct extensive experiments on different network architectures, common explainability methods, and several image datasets to demonstrate the benefits of the proposed measures. In comparison to ours, popular fidelity measures are not sufficient to guarantee trustworthy explanations. Finally, we found that 1-Lipschitz networks produce explanations with higher MeGe and ReCo than common neural networks while reaching similar accuracy. This suggests that 1-Lipschitz networks are a relevant direction towards predictors that are more explainable and trustworthy.
****
Compensation Tracker: Reprocessing Lost Object for Multi-Object Tracking
Zhibo Zou, Junjie Huang, Ping Luo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 307-317
Tracking by detection paradigm is one of the most popular object tracking methods. However, it is very dependent on the performance of the detector. When the detector has a behavior of missing detection, the tracking result will be directly affected. In this paper, we analyze the phenomenon of the lost tracking object in real-time tracking model on MOT2020 dataset. Based on simple and traditional methods, we propose a compensation tracker to further alleviate the lost tracking problem caused by missing detection. It consists of a motion compensation module and an object selection module. The proposed method not only can re-track missing tracking objects from lost objects, but also does not require additional networks so as to maintain speed-accuracy trade-off of the real-time model. Our method only needs to be embedded into the tracker to work without re-training the network. Experiments show that the compensation tracker can efficaciously improve the performance of the model and reduce identity switches. With limited costs, the compensation tracker successfully enhances the baseline tracking performance by a large margin and reaches 66% of MOTA and 67% of IDF1 on MOT2020 dataset.
****
Mending Neural Implicit Modeling for 3D Vehicle Reconstruction in the Wild
Shivam Duggal, Zihao Wang, Wei-Chiu Ma, Sivabalan Manivasagam, Justin Liang, Shenlong Wang, Raquel Urtasun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1900-1909
Reconstructing high-quality 3D objects from sparse, partial observations from a single view is of crucial importance for various applications in computer vision, robotics, and graphics. While recent neural implicit modeling methods show promising results on synthetic or dense data, they perform poorly on sparse and noisy real-world data. We discover that the limitations of a popular neural implicit model are due to lack of robust shape priors and lack of proper regularization. In this work, we demonstrate high-quality in-the-wild shape reconstruction using: (i) a deep encoder as a robust-initializer of the shape latent-code; (ii) regularized test-time optimization of the latent-code; (iii) a deep discriminator as a learned high-dimensional shape prior; (iv) a novel curriculum learning strategy that allows the model to learn shape priors on synthetic data and smoothly transfer them to sparse real-world data. Our approach better captures the global structure, performs well on occluded and sparse observations, and registers well with the ground-truth shape. We demonstrate superior performance over state-of-the-art 3D object reconstruction methods on two real-world datasets.
****
Hyperspectral Image Super-Resolution With RGB Image Super-Resolution as an Auxil

iary Task

Ke Li, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3193-3202

This work studies Hyperspectral image (HSI) super-resolution (SR). HSI SR is characterized by high-dimensional data and a limited amount of training exam-ples. This raises challenges for training deep neural net-works that are known to be data hungry. This work ad-dresses this issue with two contributions. First, we observethat HSI SR and RGB image SR are correlated and developa novel multi-tasking network to train them jointly so thatthe auxiliary task RGB image SR can provide additionalsupervision and regulate the network training. Second,we extend the network to a semi-supervised setting so thatit can learn from datasets containing only low-resolutionHSIs. With these contributions, our method is able to learnhyperspectral image super-resolution from heterogeneousdatasets and lifts the requirement for having a large amountof HD HSI training samples. Extensive experiments onthree standard datasets show that our method outperformsexisting methods significantly and underpin the relevance ofour contributions.

****

Robust Lane Detection via Expanded Self Attention

Minhyeok Lee, Junhyeop Lee, Dogyoon Lee, Woojin Kim, Sangwon Hwang, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 533-542

The image-based lane detection algorithm is one of the key technologies in autonomous vehicles. Modern deep learning methods achieve high performance in lane detection, but it is still difficult to accurately detect lanes in challenging situations such as congested roads and extreme lighting conditions. To be robust on these challenging situations, it is important to extract global contextual information even from limited visual cues. In this paper, we propose a simple but powerful self-attention mechanism optimized for lane detection called the Expanded Self Attention (ESA) module. Inspired by the simple geometric structure of lanes, the proposed method predicts the confidence of a lane along the vertical and horizontal directions in an image. The prediction of the confidence enables estimating occluded locations by extracting global contextual information. ESA module can be easily implemented and applied to any encoder-decoder-based model without increasing the inference time. The performance of our method is evaluated on three popular lane detection benchmarks (TuSimple, CULane and BDD100K). We achieve state-of-the-art performance in CULane and BDD100K and distinct improvement on TuSimple dataset. The experimental results show that our approach is robust to occlusion and extreme lighting conditions.

****

Preventing Catastrophic Forgetting and Distribution Mismatch in Knowledge Distillation via Synthetic Data

Kuluhan Binici, Nam Trung Pham, Tulika Mitra, Karianto Leman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 663-671

With the increasing popularity of deep learning on edge devices, compressing large neural networks to meet the hardware requirements of resource-constrained devices became a significant research direction. Numerous compression methodologies are currently being used to reduce the memory sizes and energy consumption of neural networks. Knowledge distillation (KD) is among such methodologies and it functions by using data samples to transfer the knowledge captured by a large model (teacher) to a smaller one (student). However, due to various reasons, the original training data might not be accessible at the compression stage. Therefore, data-free model compression is an ongoing research problem that has been addressed by various works. In this paper, we point out that catastrophic forgetting is a problem that can potentially be observed in existing data-free distillation methods. Moreover, the sample generation strategies in some of these methods could result in a mismatch between the synthetic and real data distributions. To prevent such problems, we propose a data-free KD framework that maintains a dynamic collection of generated samples over time. Additionally, we add the constraint of matching the real data distribution in sample generation strategies that ta rget maximum information gain. Our experiments demonstrate that we can improve the accuracy of the student models obtained via KD when compared with state-of-the-art approaches on the SVHN, Fashion MNIST and CIFAR100 datasets.

****

Data Augmented 3D Semantic Scene Completion With 2D Segmentation Priors

Aloisio Dourado, Frederico Guth, Teofilo de Campos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3781-3790

Semantic scene completion (SSC) is a challenging Computer Vision task with many practical applications, from robotics to assistive computing. Its goal is to infer the 3D geometry in a field of view of a scene and the semantic labels of voxels, including occluded regions. In this work, we present SPAwN, a novel lightweight multimodal 3D deep CNN that seamlessly fuses structural data from the depth component of RGB-D images with semantic priors from a bimodal 2D segmentation network. A crucial difficulty in this field is the lack of fully labeled real-world 3D datasets which are large enough to train the current data-hungry deep 3D CNNs. In 2D computer vision tasks, many data augmentation strategies have been proposed to improve the generalization ability of CNNs. However those approaches cannot be directly applied to the RGB-D input and output volume of SSC solutions. In this paper, we introduce the use of a 3D data augmentation strategy that can be applied to multimodal SSC networks. We validate our contributions with a comprehensive and reproducible ablation study. Our solution consistently surpasses previous works with a similar level of complexity.

****

Hole-Robust Wireframe Detection

Naejin Kong, Kiwoong Park, Harshith Goka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1636-1645

"Wireframe" is a line segment based representation designed to well capture large-scale visual properties of regular, structural shaped man-made scenes surrounding us. Unlike the wireframes, conventional edges or line segments focus on all visible edges and lines without particularly distinguishing which of them are more salient to man-made structural information. Existing wireframe detection models rely on supervising the annotated data but do not explicitly pay attention to understand how to compose the structural shapes of the scene. In addition, we often face that many foreground objects occluding the background scene interfere with proper inference of the full scene structure behind them. To resolve these problems, we first time in the field, propose new conditional data generation and training that help the model understand how to ignore occlusion indicated by holes, such as foreground object regions masked out on the image. In addition, we first time combine GAN in the model to let the model better predict underlying scene structure even beyond large holes. We also introduce pseudo labeling to further enlarge the model capacity to overcome small-scale labeled data. We show qualitatively and quantitatively that our approach significantly outperforms previous works unable to handle holes, as well as improves ordinary detection without holes given.

****

MUGL: Large Scale Multi Person Conditional Action Generation With Locomotion

Shubh Maheshwari, Debtanu Gupta, Ravi Kiran Sarvadevabhatla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 257-265

We introduce MUGL, a novel deep neural model for large-scale, diverse generation of single and multi-person pose-based action sequences with locomotion. Our controllable approach enables variable-length generations customizable by action category, across more than 100 categories. To enable intra/inter-category diversity, we model the latent generative space using a Conditional Gaussian Mixture Variational Autoencoder. To enable realistic generation of actions involving locomotion, we decouple local pose and global trajectory components of the action sequence. We incorporate duration-aware feature representations to enable variable-length sequence generation. We use a hybrid pose sequence representation with 3D pose sequences sourced from videos and 3D Kinect-based sequences of NTU-RGBD-120. To enable principled comparison of generation quality, we employ suitably modi fied strong baselines during evaluation. Although smaller and simpler compared to baselines, MUGL provides better quality generations, paving the way for practical and controllable large-scale human action generation.

****

Recursive Contour-Saliency Blending Network for Accurate Salient Object Detection

Yun Yi Ke, Takahiro Tsubono; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2940-2950

Contour information plays a vital role in salient object detection. However, excessive false positives remain in predictions from existing contour-based models due to insufficient contour-saliency fusion. In this work, we designed a network for better edge quality in salient object detection. We proposed a contour-saliency blending module to exchange information between contour and saliency. We adopted recursive CNN to increase contour-saliency fusion while keeping the total trainable parameters the same. Furthermore, we designed a stage-wise feature extraction module to help the model pick up the most helpful features from previous intermediate saliency predictions. Besides, we proposed two new loss functions, namely Dual Confinement Loss and Confidence Loss, for our model to generate better boundary predictions. Evaluation results on five common benchmark datasets reveal that our model achieves competitive state-of-the-art performance.

****

Multi-Branch Neural Networks for Video Anomaly Detection in Adverse Lighting and Weather Conditions

Sam Leroux, Bo Li, Pieter Simoens; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2358-2366

Automated anomaly detection in surveillance videos has attracted much interest as it provides a scalable alternative to manual monitoring. Most existing approaches achieve good performance on clean benchmark datasets recorded in well-controlled environments. However, detecting anomalies is much more challenging in the real world. Adverse weather conditions like rain or changing brightness levels cause a significant shift in the input data distribution, which in turn can lead to the detector model incorrectly reporting high anomaly scores. Additionally, surveillance cameras are usually deployed in evolving environments such as a city street of which the appearance changes over time because of seasonal changes or roadworks. The anomaly detection model will need to be updated periodically to deal with these issues. In this paper, we introduce a multi-branch model that is equipped with a trainable preprocessing step and multiple identical branches for detecting anomalies during day and night as well as in sunny and rainy conditions. We experimentally validate our approach on a distorted version of the Avenue dataset and provide qualitative results on real-world surveillance camera data. Experimental results show that our method outperforms the existing methods in terms of detection accuracy while being faster and more robust on scenes with varying visibility.

****

Natural Language Video Moment Localization Through Query-Controlled Temporal Convolution

Lingyu Zhang, Richard J. Radke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 682-690

The goal of natural language video moment localization is to locate a short segment of a long, untrimmed video that corresponds to a description presented as natural text. The description may contain several pieces of key information, including subjects/objects, sequential actions, and locations. Here, we propose a novel video moment localization framework based on the convolutional response between multimodal signals, i.e., the video sequence, the text query, and subtitles for the video if they are available. We emphasize the effect of the language sequence as a query about the video content, by converting the query sentence into a boundary detector with a filter kernel size and stride. We convolve the video sequence with the query detector to locate the start and end boundaries of the target video segment. When subtitles are available, we blend the boundary heatmaps from the visual and subtitle branches together using an LSTM to capture asynchr onous dependencies across two modalities in the video. We perform extensive experiments on the TVR, Charades-STA, and TACoS benchmark datasets, demonstrating that our model achieves state-of-the-art results on all three.

****

WEPDTOF: A Dataset and Benchmark Algorithms for In-the-Wild People Detection and Tracking From Overhead Fisheye Cameras

Ozan Tezcan, Zhihao Duan, Mertcan Cokbas, Prakash Ishwar, Janusz Konrad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 503-512

Owing to their large field of view, overhead fisheye cameras are becoming a surveillance modality of choice for large indoor spaces. However, traditional people detection and tracking algorithms developed for side-mounted, rectilinear-lens cameras do not work well on images from overhead fisheye cameras due to their viewpoint and unique optics. While several people-detection algorithms have been recently developed for such cameras, they have all been tested on datasets consisting of "staged" recordings with a limited variety of people, scenes and challenges. Clearly, the performance of these algorithms "in the wild", i.e., on recordings with real-world challenges, remains unknown. In this paper, we introduce a new benchmark dataset of in-the-Wild Events for People Detection and Tracking from Overhead Fisheye cameras (WEPDTOF). The dataset features 14 YouTube videos captured in a wide range of scenes, 188 distinct person identities consistently labeled across time, and real-world challenges such as extreme occlusions and camouflage. Also, we propose 3 spatio-temporal extensions of a state-of-the-art people-detection algorithm to enhance the coherence of detections across time. Compared to top-performing algorithms, that are purely spatial, the new algorithms offer a significant performance improvement on the new dataset. Finally, we compare the people tracking performance of these algorithms on WEPDTOF.

****

Seeing Implicit Neural Representations As Fourier Series

Nuri Benbarka, Timon Höfer, Hamd ul-Moqeet Riaz, Andreas Zell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2041-2050

Implicit Neural Representations (INR) use multilayer perceptrons to represent high-frequency functions in low-dimensional problem domains. Recently these representations achieved state-of-the-art results on tasks related to complex 3D objects and scenes. A core problem is the representation of highly detailed signals, which is tackled using networks with periodic activation functions (SIRENs) or applying Fourier mappings to the input. This work analyzes the connection between the two methods and shows that a Fourier mapped perceptron is structurally like one hidden layer SIREN. Furthermore, we identify the relationship between the previously proposed Fourier mapping and the general d-dimensional Fourier series, leading to an integer lattice mapping. Moreover, we modify a progressive training strategy to work on arbitrary Fourier mappings and show that it improves the generalization of the interpolation task. Lastly, we compare the different mappings on the image regression and novel view synthesis tasks. We confirm the previous finding that the main contributor to the mapping performance is the size of the embedding and standard deviation of its elements.

****

Human-Aided Saliency Maps Improve Generalization of Deep Learning

Aidan Boyd, Kevin W. Bowyer, Adam Czajka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2735-2744

Deep learning has driven remarkable accuracy increases in many computer vision problems. One ongoing challenge is how to achieve the greatest accuracy in cases where training data is limited. A second ongoing challenge is that trained models oftentimes do not generalize well even to new data that is subjectively similar to the training set. We address these challenges in a novel way, with the first-ever (to our knowledge) exploration of encoding human judgement about salient regions of images into the training data. We compare the accuracy and generalization of a state-of-the-art deep learning algorithm for a difficult problem in biometric presentation attack detection when trained on (a) original images with t ypical data augmentations, and (b) the same original images transformed to encode human judgement about salient image regions. The latter approach results in models that achieve higher accuracy and better generalization, decreasing the error of the LivDet-Iris 2020 winner from 29.78% to 16.37%, and achieving impressive generalization in a leave-one-attack-type-out evaluation scenario. This work opens a new area of study for how to embed human intelligence into training strategies for deep learning to achieve high accuracy and generalization in cases of limited training data.

****

Learning To Reconstruct 3D Non-Cuboid Room Layout From a Single RGB Image

Cheng Yang, Jia Zheng, Xili Dai, Rui Tang, Yi Ma, Xiaojun Yuan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2534-2543

Single-image room layout reconstruction aims to reconstruct the enclosed 3D structure of a room from a single image. Most previous work relies on the cuboid shape prior. This paper considers a more general indoor assumption, i.e., the room layout consists of a single ceiling, a single floor, and several vertical walls. To this end, we first employ Convolutional Neural Networks to detect planes and vertical lines between adjacent walls. Meanwhile, estimating the 3D parameters for each plane. Then, a simple yet effective geometric reasoning method is adopted to achieve room layout reconstruction. Furthermore, we optimize the 3D plane parameters to reconstruct a geometrically consistent room layout between planes and lines. The experimental results on public datasets validate the effectiveness and efficiency of our method.

****

Channel Pruning via Lookahead Search Guided Reinforcement Learning

Zi Wang, Chengcheng Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2029-2040

Channel pruning has become an effective yet still challenging approach to achieve compact neural networks. It aims to prune the optimal set of filters whose removal results in minimal performance degradation of the slimmed network. Due to the prohibitively vast search space of filter combinations, existing approaches usually use various criteria to estimate the filter importance while sacrificing some precision. Here we present a new approach to optimizing the filter selection in channel pruning with lookahead search guided reinforcement learning (RL). A neural network that takes as input filter-related features is trained with RL to prune the optimal sequence of filters and maximize the performance of the remaining network. In addition, we employ Monte Carlo tree search (MCTS) to provide a lookahead search for filter selection, which increases the sample efficiency for the RL training. Experiments on MNIST, CIFAR-10, and ILSVRC-2012 validate the effectiveness of our approach compared to both traditional and automated existing channel pruning approaches.

****

Self-Supervised Shape Alignment for Sports Field Registration

Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, Parthipan Siva; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 287-296

This paper presents an end-to-end self-supervised learning approach for cross-modality image registration and homography estimation, with a particular emphasis on registering sports field templates onto broadcast videos as a practical application. Rather then using any pairwise labelled data for training, we propose a self-supervised data mining method to train the registration network with a natural image and its edge map. Using an iterative estimation process controlled by a score regression network (SRN) to measure the registration error, the network can learn to estimate any homography transformation regardless of how misaligned the image and the template is. We further show the benefits of using pretrained weights to finetune the network for sports field calibration with few training data. We demonstrate the effectiveness of our proposed method by applying it to real-world sports broadcast videos where we achieve state-of-the-art results and real-time processing.

****

Intelligent Camera Selection Decisions for Target Tracking in a Camera Network

Anil Sharma, Saket Anand, Sanjit K Kaul; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3388-3397

Camera Selection Decisions (CSD) are highly useful for several applications in a multi-camera network. For example, CSD benefit multi-camera target tracking by reducing the number of candidate cameras to look for the target's next location. The correct candidate cameras, decreases the number of false Re-ID queries as well as the computation time. Also, in multi-camera trajectory forecasting (MCTF) to predict where a person will re-appear in the camera network along with the transition time. These applications require a large amount of annotated data for training. In this paper, we use state-representation learning with a reinforcement learning based policy to effectively and efficiently make camera selection decisions. We further demonstrate that by using learned state representations, as opposed to hand-crafted state variables, we are able to achieve state-of-the-art results on camera selection, while reducing the training time for the RL policy. Along with this, we use a reward function that helps to reduce the amount of supervision in training the policy in a semi-supervised way. We report our results on four datasets: NLPR-MCT, DukeMTMC, CityFlow, and WNMF dataset. We show that an RL policy reduces unnecessary Re-ID queries and therefore the false alarms, scales well to larger camera networks, and is target-agnostic.

****

Modeling Aleatoric Uncertainty for Camouflaged Object Detection

Jiawei Liu, Jing Zhang, Nick Barnes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1445-1454

Aleatoric uncertainty captures noise within the observations. For camouflaged object detection, due to similar appearance of the camouflaged foreground and the background, it's difficult to obtain highly accurate annotations, especially annotations around object boundaries. We argue that training directly with the noisy camouflage map may lead to a model of poor generalization ability. In this paper, we introduce an explicitly aleatoric uncertainty estimation technique to represent predictive uncertainty due to noisy labeling. Specifically, we present a confidence-aware camouflaged object detection (COD) framework using dynamic supervision to produce both an accurate camouflage map and a reliable aleatoric uncertainty. Different from existing techniques that produce deterministic prediction following the point estimation pipeline, our framework formalises aleatoric uncertainty as probability distribution over model output and the input image. We claim that, once trained, our confidence estimation network can evaluate the pixel-wise accuracy of the prediction without relying on the ground truth camouflage map. Extensive results illustrate the superior performance of the proposed model in explaining the camouflage prediction. Our codes are available at https://github.com/Carlisle-Liu/OCENet

****

Cross-Modal Adversarial Reprogramming

Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, Julian McAuley; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2427-2435

With the abundance of large-scale deep learning models, it has become possible to repurpose pre-trained networks for new tasks. Recent works on adversarial reprogramming have shown that it is possible to repurpose neural networks for alternate tasks without modifying the network architecture or parameters. However these works only consider original and target tasks within the same data domain. In this work, we broaden the scope of adversarial reprogramming beyond the data modality of the original task. We analyze the feasibility of adversarially repurposing image classification neural networks for Natural Language Processing (NLP) and other sequence classification tasks. We design an efficient adversarial program that maps a sequence of discrete tokens into an image which can be classified to the desired class by an image classification model. We demonstrate that by using highly efficient adversarial programs, we can reprogram image classifiers to achieve competitive performance on a variety of text and sequence classificati on benchmarks without retraining the network.

****

Learning From the CNN-Based Compressed Domain

Zhenzhen Wang, Minghai Qin, Yen-Kuang Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3582-3590

Images are transmitted or stored in their compressed form and most of the AI tasks are performed from the reconstructed domain. Convolutional neural network (CNN)-based image compression and reconstruction is growing rapidly and it achieves or surpasses the state-of-the-art heuristic image compression methods, such as JPEG or BPG. A major limitation of the application of CNN-based image compression is on the computation complexity during compression and reconstruction. Therefore, learning from the compressed domain is desirable to avoid the computation and latency caused by reconstruction. In this paper, we show that learning from the compressed domain can achieve comparative or even better accuracy than from the reconstructed domain. At a high compression rate of 0.098 bpp, for example, the proposed compression-learning system has over 3% absolute accuracy boost over the traditional compression-reconstruction-learning flow. The improvement is achieved by optimizing the compression-learning system targeting original-sized instead of standardized (e.g., 224x224) images, which is crucial in practice since real-world images into the system have different sizes. We also propose an efficient model-free entropy estimation method and a criterion to learn from a selected subset of features in the compressed domain to further reduce the transmission and computation cost without accuracy degradation.

****

Adversarial Branch Architecture Search for Unsupervised Domain Adaptation

Luca Robbiano, Muhammad Rameez Ur Rahman, Fabio Galasso, Barbara Caputo, Fabio Maria Carlucci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2918-2928

Unsupervised Domain Adaptation (UDA) is a key issue in visual recognition, as it allows to bridge different visual domains enabling robust performances in the real world. To date, all proposed approaches rely on human expertise to manually adapt a given UDA method (e.g. DANN) to a specific backbone architecture (e.g. ResNet). This dependency on handcrafted designs limits the applicability of a given approach in time, as old methods need to be constantly adapted to novel backbones. Existing Neural Architecture Search (NAS) approaches cannot be directly applied to mitigate this issue, as they rely on labels that are not available in the UDA setting. Furthermore, most NAS methods search for full architectures, which precludes the use of pre-trained models, essential in a vast range of UDA settings for reaching SOTA results. To the best of our knowledge, no prior work has addressed these aspects in the context of NAS for UDA. Here we tackle both aspects with an Adversarial Branch Architecture Search for UDA (ABAS): i. we address the lack of target labels by a novel data-driven ensemble approach for model selection; and ii. we search for an auxiliary adversarial branch, attached to a pre-trained backbone, which drives the domain alignment. We extensively validate ABAS to improve two modern UDA techniques, DANN and ALDA, on three standard visual recognition datasets (Office31, Office-Home and PACS). In all cases, ABAS robustly finds the adversarial branch architectures and parameters which yield best performances. https://github.com/lr94/abas

****

Perceptual Consistency in Video Segmentation

Yizhe Zhang, Shubhankar Borse, Hong Cai, Ying Wang, Ning Bi, Xiaoyun Jiang, Fatih Porikli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2564-2573

In this paper, we present a novel perceptual consistency perspective on video semantic segmentation, which can capture both temporal consistency and pixel-wise correctness. Given two nearby video frames, perceptual consistency measures how much the segmentation decisions agree with the pixel correspondences obtained via matching general perceptual features. More specifically, for each pixel in one frame, we find the most perceptually correlated pixel in the other frame. Our intuition is that such a pair of pixels are highly likely to belong to the same c lass. Next, we assess how much the segmentation agrees with such perceptual correspondences, based on which we derive the perceptual consistency of the segmentation maps across these two frames. Utilizing perceptual consistency, we can evaluate the temporal consistency of video segmentation by measuring the perceptual consistency over consecutive pairs of segmentation maps in a video. Furthermore, given a sparsely labeled test video, perceptual consistency can be utilized to aid with predicting the pixel-wise correctness of the segmentation on an unlabeled frame. More specifically, by measuring the perceptual consistency between the predicted segmentation and the available ground truth on a nearby frame and combining it with the segmentation confidence, we can accurately assess the classification correctness on each pixel. Our experiments show that the proposed perceptual consistency can more accurately evaluate the temporal consistency of video segmentation as compared to flow-based measures. Furthermore, it can help more confidently predict segmentation accuracy on unlabeled test frames, as compared to using classification confidence alone. Finally, our proposed measure can be used as a regularizer during the training of segmentation models, which leads to more temporally consistent video segmentation while maintaining accuracy.

****

Controlled GAN-Based Creature Synthesis via a Challenging Game Art Dataset - Addressing the Noise-Latent Trade-Off

Vaibhav Vavilala, David Forsyth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3892-3901

The state-of-the-art StyleGAN2 network supports powerful methods to create and edit art, including generating random images, finding images "like" some query, and modifying content or style. Further, recent advancements enable training with small datasets. We apply these methods to synthesize card art, by training on a novel Yu-Gi-Oh dataset. While noise inputs to StyleGAN2 are essential for good synthesis, we find that coarse-scale noise interferes with latent variables on this dataset because both control long-scale image effects. We observe over-aggressive variation in art with changes in noise and weak content control via latent variable edits. Here, we demonstrate that training a modified StyleGAN2, where coarse-scale noise is suppressed, removes these unwanted effects. We obtain a superior FID; changes in noise result in local exploration of style; and identity control is markedly improved. These results and analysis lead towards a GAN-assisted art synthesis tool for digital artists of all skill levels, which can be used in film, games, or any creative industry for artistic ideation.

****

Spatiotemporal Initialization for 3D CNNs With Generated Motion Patterns

Hirokatsu Kataoka, Kensho Hara, Ryusuke Hayashi, Eisuke Yamagata, Nakamasa Inoue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1279-1288

The paper proposes a framework of Formula-Driven Supervised Learning (FDSL) for spatiotemporal initialization. Our FDSL approach enables to automatically and simultaneously generate motion patterns and their video labels with a simple formula which is based on Perlin noise. We designed a dataset of generated motion patterns adequate for the 3D CNNs to learn a better basis set of natural videos. The constructed Video Perlin Noise (VPN) dataset can be applied to initialize a model before pre-training with large-scale video datasets such as Kinetics-400/700, to enhance target task performance. Our spatiotemporal initialization with VPN dataset (VPN initialization) outperforms the previous initialization method with the inflated 3D ConvNet (I3D) using 2D ImageNet dataset. Our proposed method increased the top-1 video-level accuracy of Kinetics-400 pre-trained model on Kinetics-400, UCF-101, HMDB-51, ActivityNet datasets. Especially, the proposed method increased the performance rate of Kinetics-400 pre-trained model by 10.3 pt on ActivityNet. We also report that the relative performance improvements from the baseline are greater in 3D CNNs rather than other models.

****

Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images

Shasvat Desai, Debasmita Ghose; Proceedings of the IEEE/CVF Winter Conference on

Applications of Computer Vision (WACV), 2022, pp. 553-563

Remote sensing data is crucial for applications ranging from monitoring forest fires and deforestation to tracking urbanization. Most of these tasks require dense pixel-level annotations for the model to parse visual information from limited labeled data available for these satellite images. Due to the dearth of high-quality labeled training data in this domain, there is a need to focus on semi-supervised techniques. These techniques generate pseudo-labels from a small set of labeled examples which are used to augment the labeled training set. This makes it necessary to have a highly representative and diverse labeled training set. Therefore, we propose to use an active learning-based sampling strategy to select a highly representative set of labeled training data. We demonstrate our proposed method's effectiveness on two existing semantic segmentation datasets containing satellite images: UC Merced Land Use Classification Dataset and DeepGlobe Land Cover Classification Dataset. We report a 27% improvement in mIoU with as little as 2% labeled data using active learning sampling strategies over randomly sampling the small set of labeled training data.

****

Multi-Stream Dynamic Video Summarization

Mohamed Elfeki, Liqiang Wang, Ali Borji; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 339-349

With vast amounts of video content being uploaded to the Internet every minute, video summarization becomes critical for efficient browsing, searching, and indexing of visual content. Nonetheless, the spread of social and egocentric cameras creates an abundance of sparse scenarios captured by several devices, and ultimately required to be jointly summarized. In this paper, we discuss the problem of summarizing videos recorded independently by several dynamic cameras that intermittently share the field of view. We present a robust framework that (a) identifies a diverse set of important events among moving cameras that often are not capturing the same scene, and (b) selects the most representative view(s) at each event to be included in a universal summary. Due to the lack of an applicable alternative, we collected a new multi-view egocentric dataset, Multi-Ego. Our dataset is recorded simultaneously by three cameras, covering a wide variety of real-life scenarios. The footage is annotated by multiple individuals under various summarization configurations, with a consensus analysis ensuring a reliable ground truth. We conduct extensive experiments on the compiled dataset in addition to three other standard benchmarks that show the robustness and the advantage of our approach in both supervised and unsupervised settings. Additionally, we show that our approach learns collectively from data of varied number-of-views and orthogonal to other summarization methods, deeming it scalable and generic. Our materials will be made publicly available.

****

Multi-Domain Semantic Segmentation With Overlapping Labels

Petra Bevandi■, Marin Orši■, Ivan Grubiši■, Josip Šari■, Siniša Šegvi■; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2615-2624

Deep supervised models have an unprecedented capacity to absorb large quantities of training data. Hence, training on many datasets becomes a method of choice towards graceful degradation in unusual scenes. Unfortunately, different datasets often use incompatible labels. For instance, the Cityscapes road class subsumes all driving surfaces, while Vistas defines separate classes for road markings, manholes etc. We address this challenge by proposing a principled method for seamless learning on datasets with overlapping classes based on partial labels and probabilistic loss. Our method achieves competitive within-dataset and cross-dataset generalization, as well as ability to learn visual concepts which are not separately labeled in any of the training datasets. Experiments reveal competitive or state-of-the-art performance on two multi-domain dataset collections and on the WildDash 2 benchmark.

****

Ortho-Shot: Low Displacement Rank Regularization With Data Augmentation for Few-Shot Learning

Uche Osahor, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2200-2209

In few-shot classification, the primary goal is to learn representations from a few samples that generalize well for novel classes. In this paper, we propose an efficient low displacement rank (LDR) regularization strategy termed Ortho-Shot; a technique that imposes orthogonal regularization on the convolutional layers of a few-shot classifier, which is based on the doubly-block toeplitz (DBT) matrix structure. The regularized convolutional layers of the few-shot classifier enhances model generalization and intra-class feature embeddings that are crucial for few-shot learning. Overfitting is a typical issue for few-shot models, the lack of data diversity inhibits proper model inference which weakens the classification accuracy of few-shot learners to novel classes. In this regard, we broke down the pipeline of the few-shot classifier and established that the support, query and task data augmentation collectively alleviates overfitting in networks. With compelling results, we demonstrated that combining a DBT-based low-rank orthogonal regularizer with data augmentation strategies, significantly boosts the performance of a few-shot classifier. We perform our experiments on the miniImagenet, CIFAR-FS and Stanford datasets with performance values of about 5% when compared to state-of-the-art.

****

Hierarchical Modeling for Task Recognition and Action Segmentation in Weakly-Labeled Instructional Videos

Reza Ghoddoosian, Saif Sayed, Vassilis Athitsos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1922-1932

This paper focuses on task recognition and action segmentation in weakly-labeled instructional videos, where only the ordered sequence of video-level actions is available during training. We propose a two-stream framework, which exploits semantic and temporal hierarchies to recognize top-level tasks in instructional videos. Further, we present a novel top-down weakly-supervised action segmentation approach, where the predicted task is used to constrain the inference of fine-grained action sequences. Experimental results on the popular Breakfast and Cooking 2 datasets show that our two-stream hierarchical task modeling significantly outperforms existing methods in top-level task recognition for all datasets and metrics. Additionally, using our task recognition framework in the proposed top-down action segmentation approach consistently improves the state of the art, while also reducing segmentation inference time by 80-90 percent.

****

Learning Temporal Video Procedure Segmentation From an Automatically Collected Large Dataset

Lei Ji, Chenfei Wu, Daisy Zhou, Kun Yan, Edward Cui, Xilin Chen, Nan Duan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1506-1515

Temporal Video Segmentation (TVS) is a fundamental video understanding task and has been widely researched in recent years. There are two subtasks of TVS: Video Action Segmentation (VAS) and Video Procedure Segmentation (VPS): VAS aims to recognize what actions happen inside the video while VPS aims to segment the video into a sequence of video clips as a procedure. The VAS task inevitably relies on pre-defined action labels and is thus hard to scale to various open-domain videos. To overcome this limitation, the VPS task tries to divide a video into several category-independent procedure segments. However, the existing dataset for the VPS task is small (2k videos) and lacks diversity (only cooking domain). To tackle these problems, we collect a large and diverse dataset called TIPS, specifically for the VPS task. TIPS contains 63k videos including more than 300k procedure segments from instructional videos on YouTube, which covers plenty of how-to areas such as cooking, health, beauty, parenting, gardening, etc. We then propose a multi-modal Transformer with Gaussian Boundary Detection (MT-GBD) model for VPS, with the backbone of the Transformer and Convolution. Furthermore, we propose a new EIOU metric for the VPS task, which helps better evaluate VPS quality in a more comprehensive way. Experimental results show the effectiveness of our proposed model and metric.

****

DAD: Data-Free Adversarial Defense at Test Time

Gaurav Kumar Nayak, Ruchit Rawal, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3562-3571

Deep models are highly susceptible to adversarial attacks. Such attacks are carefully crafted imperceptible noises that can fool the network and can cause severe consequences when deployed. To encounter them, the model requires training data for adversarial training or explicit regularization-based techniques. However, privacy has become an important concern, restricting access to only trained models but not the training data (e.g. biometric data). Also, data curation is expensive and companies may have proprietary rights over it. To handle such situations, we propose a completely novel problem of "test-time adversarial defense in absence of training data and even their statistics". We solve it in two stages: a) detection and b) correction of adversarial samples. Our adversarial sample detection framework is initially trained on arbitrary data and is subsequently adapted to the unlabelled test data through unsupervised domain adaptation. We further correct the predictions on detected adversarial samples by transforming them in Fourier domain and obtaining their low frequency component at our proposed suitable radius for model prediction. We demonstrate the efficacy of our proposed technique via extensive experiments against several adversarial attacks and for different model architectures and datasets. For a non-robust Resnet-18 model pretrained on CIFAR-10, our detection method correctly identifies 91.42% adversaries. Also, we significantly improve the adversarial accuracy from 0% to 37.37% with a minimal drop of 0.02% in clean accuracy on state-of-the-art "Auto Attack" without having to retrain the model.

****

Geometry-Inspired Top-K Adversarial Perturbations

Nurislam Tursynbek, Aleksandr Petiushko, Ivan Oseledets; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3398-3407

The brittleness of deep image classifiers to small adver-sarial input perturbations has been extensively studied inthe last several years. However, the main objective of ex-isting perturbations is primarily limited to change the cor-rectly predicted Top-1class by an incorrect one, which doesnot intend to change the Top-kprediction. In many digi-tal real-world scenarios Top-kprediction is more relevant.In this work, we propose a fast and accurate method ofcomputing Top-kadversarial examples as a simple multi-objective optimization. We demonstrate its efficacy andperformance by comparing it to other adversarial examplecrafting techniques. Moreover, based on this method, wepropose Top-kUniversal Adversarial Perturbations, image-agnostic tiny perturbations that cause the true class to beabsent among the Top-kprediction for the majority of nat-ural images. We experimentally show that our approachoutperforms baseline methods and even improves existingtechniques of finding Universal Adversarial Perturbations.

****

Shape-Coded ArUco: Fiducial Marker for Bridging 2D and 3D Modalities

Lilika Makabe, Hiroaki Santo, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2655-2664

We introduce a fiducial marker for the registration of two-dimensional (2D) images and untextured three-dimensional (3D) shapes that are recorded by commodity laser scanners. Specifically, we design a 3D-version of the ArUco marker that retains exactly the same appearance as its 2D counterpart from any viewpoint above the marker but contains shape information. The shape-coded ArUco can naturally work with off-the-shelf ArUco marker detectors in the 2D image domain. For detection in the 3D domain, we develop a method for detecting the marker in an untextured 3D point cloud. Experiments demonstrate accurate 2D-3D registration using our shape-coded ArUco markers in comparison to baseline methods.

****

Few-Shot Object Detection by Attending to Per-Sample-Prototype

Hojun Lee, Myunggi Lee, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2445-2454

Few-shot object detection aims to detect instances of specific categories in a query image with only a handful of support samples. Although this takes less effort than obtaining enough annotated images for supervised object detection, it results in a far inferior performance compared to the conventional object detection methods. In this paper, we propose a meta-learning-based approach that considers the unique characteristics of each support sample. Rather than simply averaging the information of the support samples to generate a single prototype per category, our method can better utilize the information of each support sample by treating each support sample as an individual prototype. Specifically, we introduce two types of attention mechanisms for aggregating the query and support feature maps. The first is to refine the information of few-shot samples by extracting shared information between the support samples through attention. Second, each support sample is used as a class code to leverage the information by comparing similarities between each support feature and query features. Our proposed method is complementary to the previous methods, making it easy to plug and play for further improvement. We have evaluated our method on PASCAL VOC and COCO benchmarks, and the results verify the effectiveness of our method. In particular, the advantages of our method is maximized when there is more diversity among support data.

****

How and What To Learn: Taxonomizing Self-Supervised Learning for 3D Action Recognition

Amor Ben Tanfous, Aimen Zerroug, Drew Linsley, Thomas Serre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2696-2705

There are two competing standards for self-supervised learning in action recognition from 3D skeletons. Su et al., 2020 used an auto-encoder architecture and an image reconstruction objective function to achieve state-of-the-art performance on the NTU60 C-View benchmark. Rao et al., 2020 used Contrastive learning in the latent space to achieve state-of-the-art performance on the NTU60 C-Sub benchmark. Here, we reconcile these disparate approaches by developing a taxonomy of self-supervised learning for action recognition. We observe that leading approaches generally use one of two types of objective functions: those that seek to reconstruct the input from a latent representation ("Attractive" learning) versus those that also try to maximize the representations distinctiveness ("Contrastive" learning). Independently, leading approaches also differ in how they implement these objective functions: there are those that optimize representations in the decoder output space and those which optimize representations in the network's latent space (encoder output). We find that combining these approaches leads to larger gains in performance and tolerance to transformation than is achievable by any individual method, leading to state-of-the-art performance on three standard action recognition datasets. We include links to our code and data.

****

Interpretable Semantic Photo Geolocation

Jonas Theiner, Eric Müller-Budack, Ralph Ewerth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 750-760

Planet-scale photo geolocalization is the complex task of estimating the location depicted in an image solely based on its visual content. Due to the success of convolutional neural networks (CNNs), current approaches achieve super-human performance. However, previous work has exclusively focused on optimizing geolocalization accuracy. Due to the black-box property of deep learning systems, their predictions are difficult to validate for humans. State-of-the-art methods treat the task as a classification problem, where the choice of the classes, that is the partitioning of the world map, is crucial for the performance. In this paper, we present two contributions to improve the interpretability of a geolocalization model: (1) We propose a novel semantic partitioning method which intuitively leads to an improved understanding of the predictions, while achieving state-of-the-art results for geolocational accuracy on benchmark test sets; (2) We intro duce a metric to assess the importance of semantic visual concepts for a certain prediction to provide additional interpretable information, which allows for a large-scale analysis of already trained models. Source code and dataset are publicly available.

****

Leaky Gated Cross-Attention for Weakly Supervised Multi-Modal Temporal Action Localization

Jun-Tae Lee, Sungrack Yun, Mihir Jain; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3213-3222

As multiple modalities sometimes have a weak complementary relationship, multi-modal fusion is not always beneficial for weakly supervised action localization. Hence, to attain the adaptive multi-modal fusion, we propose a leaky gated cross-attention mechanism. In our work, we take the multi-stage cross-attention as the baseline fusion module to obtain multi-modal features. Then, for the stages of each modality, we design gates to decide the dependency on the other modality. For each input frame, if two modalities have a strong complementary relationship, the gate selects the cross-attended feature, otherwise the non-attended feature. Also, the proposed gate allows the non-selected feature to escape through it with a small intensity, we call it leaky gate. This leaky feature makes effective regularization of the selected major feature. Therefore, our leaky gating makes cross-attention more adaptable and robust even when the modalities have a weak complementary relationship. The proposed leaky gated cross-attention provides a modality fusion module that is generally compatible with various temporal action localization methods. To show its effectiveness, we do extensive experimental analysis and apply the proposed method to boost the performance of the state-of-the-art methods on two benchmark datasets (ActivityNet1.2 and THUMOS14).

****

Complete Face Recovery GAN: Unsupervised Joint Face Rotation and De-Occlusion From a Single-View Image

Yeong-Joon Ju, Gun-Hee Lee, Jung-Ho Hong, Seong-Whan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3711-3721

Although various face-related tasks have significantly advanced in recent years, occlusion and extreme pose still impede the achievement of higher performance. Existing face rotation or de-occlusion methods only have emphasized the aspect of each problem. In addition, the lack of high-quality paired data remains an obstacle for both methods. In this work, we present a self-supervision strategy called Swap-R&R to overcome the lack of ground-truth in a fully unsupervised manner for joint face rotation and de-occlusion. To generate an input pair for self-supervision, we transfer the occlusion from a face in an image to an estimated 3D face and create a damaged face image, as if rotated from a different pose by rotating twice with the roughly de-occluded face. Furthermore, we propose Complete Face Recovery GAN (CFR-GAN) to restore the collapsed textures and disappeared occlusion areas by leveraging the structural and textural differences between two rendered images. Unlike previous works, which have selected occlusion-free images to obtain ground-truths, our approach does not require human intervention and paired data. We show that our proposed method can generate a de-occluded frontal face image from an occluded profile face image. Moreover, extensive experiments demonstrate that our approach can boost the performance of facial recognition and facial expression recognition. The code is publicly available.

****

A Semi-Supervised Generalized VAE Framework for Abnormality Detection Using One-Class Classification

Renuka Sharma, Satvik Mashkaria, Suyash P. Awate; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 595-603

Anomaly detection is a one-class classification (OCC) problem where the methods learn either a generative model of the inlier class (e.g., in the variants of kernel principal component analysis) or a decision boundary to encapsulate the inlier class (e.g., in the one-class variants of the support vector machine). Learning schemes for OCC typically rely on training data solely from the inlier class

, but some recent approaches have proposed semi-supervised extensions, e.g., variants of semi-supervised anomaly detection that also leverage a small amount of training data from outlier classes. Other recent methods extend existing principles to employ deep neural network (DNN) modeling that relies on learning (for the inlier class) either latent-space distributions or autoencoders, but not both. We propose a novel semi-supervised variational formulation, leveraging generalized-Gaussian models leading to data-adaptive, robust, and uncertainty-aware distribution modeling in both latent space and image space. For variational learning, we propose a novel reparameterization for sampling from the latent-space generalized-Gaussian to enable backpropagation-based optimization. Results on several public image sets show the benefits of our method over state of the art.

****

Mesh Convolutional Autoencoder for Semi-Regular Meshes of Different Sizes

Sara Hahner, Jochen Garcke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 885-894

The analysis of deforming 3D surface meshes is accelerated by autoencoders since the low-dimensional embeddings can be used to visualize underlying dynamics. But, state-of-the-art mesh convolutional autoencoders require a fixed connectivity of all input meshes handled by the autoencoder. This is due to either the use of spectral convolutional layers or mesh dependent pooling operations. Therefore, the types of datasets that one can study are limited and the learned knowledge cannot be transferred to other datasets that exhibit similar behavior. To address this, we transform the discretization of the surfaces to semi-regular meshes that have a locally regular connectivity and whose meshing is hierarchical. This allows us to apply the same spatial convolutional filters to the local neighborhoods and to define a pooling operator that can be applied to every semi-regular mesh. We apply the same mesh autoencoder to different datasets and our reconstruction error is more than 50% lower than the error from state-of-the-art models, which have to be trained for every mesh separately. Additionally, we visualize the underlying dynamics of unseen mesh sequences with an autoencoder trained on different classes of meshes.

****

PoP-Net: Pose Over Parts Network for Multi-Person 3D Pose Estimation From a Depth Image

Yuliang Guo, Zhong Li, Zekun Li, Xiangyu Du, Shuxue Quan, Yi Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1240-1249

In this paper, a real-time method called PoP-Net is proposed to predict multi-person 3D poses from a depth image. PoP-Net learns to predict bottom-up part representations and top-down global poses in a single shot. Specifically, a new part-level representation, called Truncated Part Displacement Field (TPDF), is introduced which enables an explicit fusion process to unify the advantages of bottom-up part detection and global pose detection. Meanwhile, an effective mode selection scheme is introduced to automatically resolve the conflicting cases between global pose and part detections. Finally, due to the lack of high-quality depth datasets for developing multi-person 3D pose estimation, we introduce Multi-Person 3D Human Pose Dataset (MP-3DHP) as a new benchmark. MP-3DHP is designed to enable effective multi-person and background data augmentation in model training, and to evaluate 3D human pose estimators under uncontrolled multi-person scenarios. We show that PoP-Net achieves the state-of-the-art results both on MP-3DHP and on the widely used ITOP dataset, and has significant advantages in efficiency for multi-person processing. MP-3DHP Dataset and the evaluation code have been made available at: https://github.com/oppo-us-research/PoP-Net.

****

Fusion Point Pruning for Optimized 2D Object Detection With Radar-Camera Fusion

Lukas Stäcker, Philipp Heidenreich, Jason Rambach, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3087-3094

Object detection is one of the most important perception tasks for advanced driver assistant systems and autonomous driving. Due to its complementary features a nd moderate cost, radar-camera fusion is of particular interest in the automotive industry but comes with the challenge of how to optimally fuse the heterogeneous data sources. To solve this for 2D object detection, we propose two new techniques to project the radar detections onto the image plane, exploiting additional uncertainty information. We also introduce a new technique called fusion point pruning, which automatically finds the best fusion points of radar and image features in the neural network architecture. These new approaches combined surpass the state of the art in 2D object detection performance for radar-camera fusion models, evaluated with the nuScenes dataset. We further find that the utilization of radar-camera fusion is especially beneficial for night scenes.

****

CoordiNet: Uncertainty-Aware Pose Regressor for Reliable Vehicle Localization

Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, Arnaud de La Fortelle; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2229-2238

In this paper, we investigate visual-based camera relocalization with neural networks for robotics and autonomous vehicles applications. Our solution is a CNN-based algorithm which predicts camera pose (3D translation and 3D rotation) directly from a single image. It also provides an uncertainty estimate of the pose. Pose and uncertainty are learned together with a single loss function and are fused at test time with an EKF. Furthermore, we propose a new fully convolutional architecture, named CoordiNet, designed to embed some of the scene geometry. Our framework outperforms comparable methods on the largest available benchmark, the Oxford RobotCar dataset, with an average error of 8 meters where previous best was 19 meters. We have also investigated the performance of our method on large scenes for real time (18 fps) vehicle localization. In this setup, structure-based methods require a large database, and we show that our proposal is a reliable alternative, achieving 29cm median error in a 1.9km loop in a busy urban area.

****

TA-Net: Topology-Aware Network for Gland Segmentation

Haotian Wang, Min Xian, Aleksandar Vakanski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1556-1564

Gland segmentation is a critical step to quantitatively assess the morphology of glands in histopathology image analysis. However, it is challenging to separate densely clustered glands accurately. Existing deep learning-based approaches attempted to use contour-based techniques to alleviate this issue but only achieved limited success. To address this challenge, we propose a novel topology-aware network (TA-Net) to accurately separate densely clustered and severely deformed glands. The proposed TA-Net has a multitask learning architecture and enhances the generalization of gland segmentation by learning shared representation from two tasks: instance segmentation and gland topology estimation. The proposed topology loss computes gland topology using gland skeletons and markers. It drives the network to generate segmentation results that comply with the true gland topology. We validate the proposed approach on the GlaS and CRAG datasets using three quantitative metrics, F1-score, object-level Dice coefficient, and object-level Hausdorff distance. Extensive experiments demonstrate that TA-Net achieves state-of-the-art performance on the two datasets. TA-Net outperforms other approaches in the presence of densely clustered glands.

****

YOLO-ReT: Towards High Accuracy Real-Time Object Detection on Edge GPUs

Prakhar Ganesh, Yao Chen, Yin Yang, Deming Chen, Marianne Winslett; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3267-3277

Performance of object detection models has been growing rapidly on two major fronts, model accuracy and efficiency. However, in order to map deep neural network (DNN) based object detection models to edge devices, one typically needs to compress such models significantly, thus compromising the model accuracy. In this paper, we propose a novel edge GPU friendly module for multi-scale feature interaction by exploiting missing combinatorial connections between various feature scales in existing state-of-the-art methods. Additionally, we propose a novel tran sfer learning backbone adoption inspired by the changing translational information flow across various tasks, designed to complement our feature interaction module and together improve both accuracy as well as execution speed on various edge GPU devices available in the market. For instance, YOLO-ReT with MobileNetV2x0.75 backbone runs real-time on Jetson Nano, and achieves 68.75 mAP on Pascal VOC and 34.91 mAP on COCO, beating its peers by 3.05 mAP and 0.91 mAP respectively, while executing faster by 3.05 FPS. Furthermore, introducing our multi-scale feature interaction module in YOLOv4-tiny and YOLOv4-tiny (3l) improves their performance to 41.5 and 48.1 mAP respectively on COCO, outperforming the original versions by 1.3 and 0.9 mAP.

****

Mixed-Dual-Head Meets Box Priors: A Robust Framework for Semi-Supervised Segmentation

Chenshu Chen, Tao Liu, Wenming Tan, Shiliang Pu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1098-1108

As it is costly to densely annotate large scale datasets for supervised semantic segmentation, extensive semi-supervised methods have been proposed. However, the accuracy, stability and flexibility of existing methods are still far from satisfactory. In this paper, we propose an effective and flexible framework for semi-supervised semantic segmentation using a small set of fully labeled images and a set of weakly labeled images with bounding box labels. In our framework, position and class priors are designed to guide the annotation network to predict accurate pseudo masks for weakly labeled images, which are used to train the segmentation network. We also propose a mixed-dual-head training method to reduce the interference of label noise while enabling the training process more stable. Experiments on PASCAL VOC 2012 show that our method achieves state-of-the-art performance and can achieve competitive results even with very few fully labeled images. Furthermore, the performance can be further boosted with extra weakly labeled images from COCO dataset.

****

Meta-Learning for Multi-Label Few-Shot Classification

Christian Simon, Piotr Koniusz, Mehrtash Harandi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3951-3960

Even with the luxury of having abundant data, multi-label classification is widely known to be a challenging task to address. This work targets the problem of multi-label meta-learning, where a model learns to predict multiple labels within a query (e.g., an image) by just observing a few supporting examples. In doing so, we first propose a benchmark for Few-Shot Learning (FSL) with multiple labels per sample. Next, we discuss and extend several solutions specifically designed to address the conventional and single-label FSL, to work in the multi-label regime. Lastly, we introduce a neural module to estimate the label count of a given sample by exploiting the relational inference. We will show empirically the benefit of the label count module, the label propagation algorithm, and the extensions of conventional FSL methods on three challenging datasets, namely MS-COCO, iMaterialist, and Open MIC. Overall, our thorough experiments suggest that the proposed label-propagation algorithm in conjunction with the neural label count module (NLC) shall be considered as the method of choice.

****

Modeling Dynamic Target Deformation in Camera Calibration

Annika Hagemann, Moritz Knorr, Christoph Stiller; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1747-1755

Most approaches to camera calibration rely on calibration targets of well-known geometry. During data acquisition, calibration target and camera system are typically moved w.r.t. each other, to allow image coverage and perspective versatility. We show that moving the target can lead to small temporary deformations of the target, which can introduce significant errors into the calibration result. While static inaccuracies of calibration targets have been addressed in previous works, to our knowledge, none of the existing approaches can capture time-varying, dynamic deformations. To achieve high-accuracy calibrations despite moving the target, we propose a way to explicitly model dynamic target deformations in ca mera calibration. This is achieved by using a low-dimensional deformation model with only few parameters per image, which can be optimized jointly with target poses and intrinsics. We demonstrate the effectiveness of modeling dynamic deformations using different calibration targets and show its significance in a structure-from-motion application.

****

MTGLS: Multi-Task Gaze Estimation With Limited Supervision

Shreya Ghosh, Munawar Hayat, Abhinav Dhall, Jarrod Knibbe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3223-3234

Robust gaze estimation is a challenging task, even for deep CNNs, due to the non-availability of large-scale labeled data. Moreover, gaze annotation is a time-consuming process and requires specialized hardware setups. We propose MTGLS: a Multi-Task Gaze estimation framework with Limited Supervision, which leverages abundantly available non-annotated facial image data. MTGLS distills knowledge from off-the-shelf facial image analysis models, and learns strong feature representations of human eyes, guided by three complementary auxiliary signals: (a) the line of sight of the pupil (i.e. pseudo-gaze) defined by the localized facial landmarks, (b) the head-pose given by Euler angles, and (c) the orientation of the eye patch (left/right eye). To overcome inherent noise in the supervisory signals, MTGLS further incorporates a noise distribution modelling approach. Our experimental results show that MTGLS learns highly generalized representations which consistently perform well on a range of datasets. Our proposed framework outperforms the unsupervised state-of-the-art on CAVE (by approx. 6.43%) and even supervised state-of-the-art methods on Gaze360 (by approx. 6.59%) datasets.

****

Online Continual Learning via Candidates Voting

Jiangpeng He, Fengqing Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3154-3163

Continual learning in online scenario aims to learn a sequence of new tasks from data stream using each data only once for training, which is more realistic than in offline mode assuming data from new task are all available. However, this problem is still under-explored for the challenging class-incremental setting in which the model classifies all classes seen so far during inference. Particularly, performance struggles with increased number of tasks or additional classes to learn for each task. In addition, most existing methods require storing original data as exemplars for knowledge replay, which may not be feasible for certain applications with limited memory budget or privacy concerns. In this work, we introduce an effective and memory-efficient method for online continual learning under class-incremental setting through candidates selection from each learned task together with prior incorporation using stored feature embeddings instead of original data as exemplars. Our proposed method implemented for image classification task achieves the best results under different benchmark datasets for online continual learning including CIFAR-10, CIFAR-100 and CORE-50 while requiring much less memory resource compared with existing works.

****

Geometry-Aware Hierarchical Bayesian Learning on Manifolds

Yonghui Fan, Yalin Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1786-1795

Bayesian learning with Gaussian processes demonstrates encouraging regression and classification performance in solving computer vision tasks. However, Bayesian methods on 3D manifold-valued vision data, such as meshes and point clouds, are seldom studied. One of the primary challenges is how to effectively and efficiently aggregate geometric features from inputs. In this paper, we propose a hierarchical Bayesian learning model to address this challenge. We implicitly introduce the geometry-awareness and the intra-kernel convolution to the kernel so that the prior becomes geometry sensitive without using any hand-crafted feature descriptors. We implement a hierarchical feature aggregation architecture by concatenating multiple Gaussian processes together. Furthermore, we incorporate the feature learning of neural networks with the feature aggregation of Bayesian model s to investigate the feasibility of jointly learning inferences on manifolds. Experimental results not only show that our method outperforms existing Bayesian methods on manifolds but also demonstrate the prospect of coupling neural networks with Bayesian learning methods

****

MaskSplit: Self-Supervised Meta-Learning for Few-Shot Semantic Segmentation

Mustafa Sercan Amac, Ahmet Sencan, Bugra Baran, Nazli Ikizler-Cinbis, Ramazan Gokberk Cinbis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1067-1077

Just like other few-shot learning problems, few-shot segmentation aims to minimize the need for manual annotation, which is particularly costly in segmentation tasks. Even though the few-shot setting reduces this cost for novel test classes, there is still a need to annotate the training data. To alleviate this need, we propose a self-supervised training approach for learning few-shot segmentation models. We first use unsupervised saliency estimation to obtain pseudo-masks on images. We then train a simple prototype based model over different splits of pseudo masks and augmentations of images. Our extensive experiments show that the proposed approach achieves promising results, highlighting the potential of self-supervised training. To the best of our knowledge this is the first work that addresses unsupervised few-shot segmentation problem on natural images.

****

Measuring Hidden Bias Within Face Recognition via Racial Phenotypes

Seyma Yucer, Furkan Tektas, Noura Al Moubayed, Toby P. Breckon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 995-1004

Recent work reports disparate performance for intersectional racial groups across face recognition tasks: face verification and identification. However, the definition of racial groups has a significant impact on the underlying findings of such racial bias analysis. Previous studies define these groups based on either demographic information (e.g. African, Asian etc.) or skin tone (e.g. lighter or darker skins). The use of such either sensitive or broad and loosely defined group definitions has disadvantages for both bias investigation and the design of subsequent counter-bias solutions. By contrast, this study introduces an alternative racial bias analysis methodology via the use of facial phenotype attributes for face recognition. We use the set of observable characteristics of an individual face where a race-related facial phenotype is hence specific to the human face and correlated to the racial profile of the subject. We propose categorical test cases to investigate the individual influence of those attributes on bias within face recognition tasks. We compare our phenotype-based grouping methodology with previous grouping strategies and show that phenotype-based groupings uncover hidden bias without exposing any potentially protected attributes. Furthermore, we contribute corresponding phenotype attribute category labels for face recognition tasks: RFW for face verification and VGGFace2 (test set) for face identification.

****

Transferable 3D Adversarial Textures Using End-to-End Optimization

Camilo Pestana, Naveed Akhtar, Nazanin Rahnavard, Mubarak Shah, Ajmal Mian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 88-97

Deep visual models are known to be vulnerable to adversarial attacks. The last few years have seen numerous techniques to compute adversarial inputs for these models. However, there are still under-explored avenues in this critical research direction. Among those is the estimation of adversarial textures for 3D models in an end-to-end optimization scheme. In this paper, we propose such a scheme to generate adversarial textures for 3D models that are highly transferable and invariant to different camera views and lighting conditions. Our method makes use of neural rendering with explicit control over the model texture and background. We ensure transferability of the adversarial textures by employing an ensemble of robust and non-robust models. Our technique utilizes 3D models as a proxy to simulate closer to real-life conditions, in contrast to conventional use of 2D i mages for adversarial attacks. We show the efficacy of our method with extensive experiments.

****

MEGAN: Memory Enhanced Graph Attention Network for Space-Time Video Super-Resolution

Chenyu You, Lianyi Han, Aosong Feng, Ruihan Zhao, Hui Tang, Wei Fan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1401-1411

Space-time video super-resolution (STVSR) aims to construct a high space-time resolution video sequence from the corresponding low-frame-rate, low-resolution video sequence. Inspired by the recent success to consider spatial-temporal information for space-time super-resolution, our main goal in this work is to take full considerations of spatial and temporal correlations within the video sequences of fast dynamic events. To this end, we propose a novel one-stage memory enhanced graph attention network (MEGAN) for space-time video super-resolution. Specifically, we build a novel long-range memory graph aggregation (LMGA) module to dynamically capture correlations along the channel dimensions of the feature maps and adaptively aggregate channel features to enhance the feature representations. We introduce a non-local residual block, which enables each channel-wise feature to attend global spatial hierarchical features. In addition, we adopt a progressive fusion module to further enhance the representation ability by extensively exploiting spatio-temporal correlations from multiple frames. Experiment results demonstrate that our method achieves better results compared with the state-of-the-art methods quantitatively and visually.

****

Single-Photon Camera Guided Extreme Dynamic Range Imaging

Yuhao Liu, Felipe Gutierrez-Barragan, Atul Ingle, Mohit Gupta, Andreas Velten; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1575-1585

Reconstruction of high-resolution extreme dynamic range images from a small number of low dynamic range (LDR) images is crucial for many computer vision applications. Current high dynamic range (HDR) cameras based on CMOS image sensor technology rely on multiexposure bracketing which suffers from motion artifacts and signal-to-noise (SNR) dip artifacts in extreme dynamic range scenes. Recently, single-photon cameras (SPCs) have been shown to achieve orders of magnitude higher dynamic range for passive imaging than conventional CMOS sensors. SPCs are becoming increasingly available commercially, even in some consumer devices. Unfortunately, current SPCs suffer from low spatial resolution. To overcome the limitations of CMOS and SPC sensors, we propose a learning-based CMOS-SPC fusion method to recover high-resolution extreme dynamic range images. We compare the performance of our method against various traditional and state-of-the-art baselines using both synthetic and experimental data. Our method outperforms these baselines, both in terms of visual quality and quantitative metrics.

****

SC-UDA: Style and Content Gaps Aware Unsupervised Domain Adaptation for Object Detection

Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, Xiang Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 382-391

Current state-of-the-art object detectors can have a significant performance drop when deployed in the wild due to domain gaps with training data. Unsupervised Domain Adaptation (UDA) is a promising approach to adapt detectors for new domains/environments without any expensive label cost. Previous mainstream UDA works for object detection usually focused on image-level and/or feature-level adaptation by using adversarial learning methods. In this work, we show that such adversarial-based methods can only reduce the domain style gap, but cannot address the domain content gap that is also important for object detectors. To overcome this limitation, we propose the SC-UDA framework to concurrently reduce both gaps: We propose fine-grained domain style transfer to reduce the style gaps with finer image details preserved for detecting small objects; Then we leverage the pse

udo-label-based self-training to reduce content gaps; To address pseudo label error accumulation during self-training, novel optimizations are proposed, including uncertainty-based pseudo labeling and imbalanced mini-batch sampling strategy. Experiment results show that our approach consistently outperforms prior stat-of-the-art methods (up to 8.6%, 2.7%, and 2.5% mAP on three UDA benchmarks).
****

Lightweight Monocular Depth With a Novel Neural Architecture Search Method
Lam Huynh, Phong Nguyen, Ji■í Matas, Esa Rahtu, Janne Heikkilä; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3643-3653

This paper presents a novel neural architecture search method, called LiDNAS, for generating lightweight monocular depth estimation models. Unlike previous neural architecture search (NAS) approaches, where finding optimized networks is computationally highly demanding, the introduced novel Assisted Tabu Search leads to efficient architecture exploration. Moreover, we construct the search space on a pre-defined backbone network to balance layer diversity and search space size. The LiDNAS method outperforms the state-of-the-art NAS approach, proposed for disparity and depth estimation, in terms of search efficiency and output model performance. The LiDNAS optimized models achieve result superior to compact depth estimation state-of-the-art on NYU-Depth-v2, KITTI, and ScanNet, while being 7%-500% more compact in size, i.e the number of model parameters.
****

Progressive Automatic Design of Search Space for One-Shot Neural Architecture Search
Xin Xia, Xuefeng Xiao, Xing Wang, Min Zheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2455-2464

Neural Architecture Search (NAS) has attracted growing interest. To reduce the search cost, recent work has explored weight sharing across models and made major progress in One-Shot NAS. However, it has been observed that a model with higher one-shot model accuracy does not necessarily perform better when stand-alone trained. To address this issue, in this paper, we propose Progressive Automatic Design of search space, named PAD-NAS. Unlike previous approaches where the same operation search space is shared by all the layers in the supernet, we formulate a progressive search strategy based on operation pruning and build a layer-wise operation search space. In this way, PAD-NAS can automatically design the operations for each layer and achieve a trade-off between search space quality and model diversity. During the search, we also take the hardware platform constraints into consideration for efficient neural network model deployment. Extensive experiments on ImageNet show that our method can achieve state-of-the-art performance.
****

Improving Fractal Pre-Training
Connor Anderson, Ryan Farrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1300-1309

The deep neural networks used in modern computer vision systems require enormous image datasets to train them. These carefully-curated datasets typically have a million or more images, across a thousand or more distinct categories. The process of creating and curating such a dataset is a monumental undertaking, demanding extensive effort and labelling expense and necessitating careful navigation of technical and social issues such as label accuracy, copyright ownership, and content bias. What if we had a way to harness the power of large image datasets but with few or none of the major issues and concerns currently faced? This paper extends the recent work of Kataoka et al. [2020], proposing an improved pre-training dataset based on dynamically-generated fractal images. Challenging issues with large-scale image datasets become points of elegance for fractal pre-training: perfect label accuracy at zero cost; no need to store/transmit large image archives; no privacy/demographic bias/concerns of inappropriate content, as no humans are pictured; limitless supply and diversity of images; and the images are free/open-source. Perhaps surprisingly, avoiding these difficulties imposes only a small penalty in performance. Leveraging a newly-proposed pre-training task--

-multi-instance prediction---our experiments demonstrate that fine-tuning a network pre-trained using fractals attains 92.7-98.1% of the accuracy of an ImageNet pre-trained network. Our code is publicly available.

****

EllipsoidNet: Ellipsoid Representation for Point Cloud Classification and Segmentation

Yecheng Lyu, Xinming Huang, Ziming Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 854-864

Point cloud patterns are hard to learn because of the implicit local geometry features among the orderless points. In recent years, point cloud representation in 2D space has attracted increasing research interest since it exposes the local geometry features in a 2D space. By projecting those points to a 2D feature map, the relationship between points is inherited in the context between pixels, which are further extracted by a 2D convolutional neural network. However, existing 2D representing methods are either accuracy limited or time-consuming. In this paper, we propose a novel 2D representation method that projects a point cloud onto an ellipsoid surface space, where local patterns are well exposed in ellipsoid-level and point-level. Additionally, a novel convolutional neural network named EllipsoidNet is proposed to utilize those features for point cloud classification and segmentation applications. The proposed methods are evaluated in ModelNet40 and ShapeNet benchmarks, where the advantages are clearly shown over existing 2D representation methods.

****

Deep Two-Stream Video Inference for Human Body Pose and Shape Estimation

Ziwen Li, Bo Xu, Han Huang, Cheng Lu, Yandong Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 430-439

Several video-based 3D pose and shape estimation algorithms have been proposed to resolve the temporal inconsistency of single-image-based counterparts. However it still remains chanllenging to have stable and accurate reconstruction. In this paper, we propose a new method Deep Two-Stream Video Inference for Human Body Pose and Shape Estimation (DTS-VIBE), to generate 3D human pose and mesh from RGB videos. We reformulate the task as a multi-modality problem that fuses RGB and optical flow for more reliable estimation. In order to fully utilize both sensory modalities (RGB or optical flow), we train a two-stream temporal network based on transformer to predict SMPL parameters. The supplementary modality, optical flow, helps to maintain temporal consistency by leveraging motion knowlege between two consecutive frames. The proposed algorithm is extensively evaluated on the Human3.6 and 3DPW datasets. The experimental results show that it outperforms other state-of-the-art methods by a significant margin.

****

M3DETR: Multi-Representation, Multi-Scale, Mutual-Relation 3D Object Detection With Transformers

Tianrui Guan, Jun Wang, Shiyi Lan, Rohan Chandra, Zuxuan Wu, Larry Davis, Dinesh Manocha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 772-782

We present a novel architecture for 3D object detection, M3DETR, which combines different point cloud representations (raw, voxels, bird-eye view) with different feature scales based on multi-scale feature pyramids. M3DETR is the first approach that unifies multiple point cloud representations, feature scales, as well as models mutual relationships between point clouds simultaneously using transformers. We perform extensive ablation experiments that highlight the benefits of fusing representation and scale, and modeling the relationships. Our method achieves state-of-the-art performance on the KITTI 3D object detection dataset and Waymo Open Dataset. Results show that M3DETR improves the baseline significantly by 1.48% mAP for all classes on Waymo Open Dataset. In particular, our approach ranks 1st on the well-known KITTI 3D Detection Benchmark for both car and cyclist classes, and ranks 1st on Waymo Open Dataset with single frame point cloud input. Our code is available at: https://github.com/rayguan97/M3DETR.

****

Tensor-Based Non-Rigid Structure From Motion

Stella Graßhof, Sami Sebastian Brandt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3011-3020

In this work we present a method that combines tensor-based face modelling and analysis and non-rigid structure-from-motion (NRSFM). The core idea is to see that the conventional tensor formulation for the face structure and expression analysis can be utilised while the structure component can be directly analysed as the non-rigid structure-from-motion problem. To the NRSFM problem part we further present a novel prior-free approach that factorises the 2D input shapes into affine projection matrices, rank-one 3D affine basis shapes, and the basis shape coefficients. The linear combination of the basis shapes thus yields the recovered 3D shapes upto an affine transformation. In contrast to most works in literature, no calibration information of the cameras or structure prior is required. Experiments on challenging face datasets show that our method, with and without the metric upgrade, is accurate and fast when compared to the state-of-the-art and is well suitable for dense reconstruction and face editing.

****

Data InStance Prior (DISP) in Generative Adversarial Networks

Puneet Mangla, Nupur Kumari, Mayank Singh, Balaji Krishnamurthy, Vineeth N. Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 451-461

Recent advances in generative adversarial networks (GANs) have shown remarkable progress in generating high-quality images. However, this gain in performance depends on the availability of a large amount of training data. In limited data regimes, training typically diverges, and therefore the generated samples are of low quality and lack diversity. Previous works have addressed training in low data setting by leveraging transfer learning and data augmentation techniques. We propose a novel transfer learning method for GANs in the limited data domain by leveraging informative data prior derived from self-supervised/supervised pre-trained networks trained on a diverse source domain. We perform experiments on several standard vision datasets using various GAN architectures (BigGAN, SNGAN, StyleGAN2) to demonstrate that the proposed method effectively transfers knowledge to domains with few target images, outperforming existing state-of-the-art techniques in terms of image quality and diversity. We also show the utility of data instance prior in large-scale unconditional image generation.

****

EdgeConv With Attention Module for Monocular Depth Estimation

Minhyeok Lee, Sangwon Hwang, Chaewon Park, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2858-2867

Monocular depth estimation is an especially important task in robotics and autonomous driving, where 3D structural information is essential. However, extreme lighting conditions and complex surface objects make it difficult to predict depth in a single image. Therefore, to generate accurate depth maps, it is important for the model to learn structural information about the scene. We propose a novel Patch-Wise EdgeConv Module (PEM) and EdgeConv Attention Module (EAM) to solve the difficulty of monocular depth estimation. The proposed modules extract structural information by learning the relationship between image patches close to each other in space using edge convolution. Our method is evaluated on two popular datasets, the NYU Depth V2 and the KITTI Eigen split, achieving state-of-the-art performance. We prove that the proposed model predicts depth robustly in challenging scenes through various comparative experiments.

****

Self-Supervised Video Representation Learning With Cross-Stream Prototypical Contrasting

Martine Toering, Ioannis Gatopoulos, Maarten Stol, Vincent Tao Hu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 108-118

Instance-level contrastive learning techniques, which rely on data augmentation and a contrastive loss function, have found great success in the domain of visual representation learning. They are not suitable for exploiting the rich dynamic al structure of video however, as operations are done on many augmented instances. In this paper we propose "Video Cross-Stream Prototypical Contrasting", a novel method which predicts consistent prototype assignments from both RGB and optical flow views, operating on sets of samples. Specifically, we alternate the optimization process; while optimizing one of the streams, all views are mapped to one set of stream prototype vectors. Each of the assignments is predicted with all views except the one matching the prediction, pushing representations closer to their assigned prototypes. As a result, more efficient video embeddings with ingrained motion information are learned, without the explicit need for optical flow computation during inference. We obtain state-of-the-art results on nearest-neighbour video retrieval and action recognition, outperforming previous best by +3.2% on UCF101 using the S3D backbone (90.5% Top-1 acc), and by +7.2% on UCF101 and +15.1% on HMDB51 using the R(2+1)D backbone.

****

To Miss-Attend Is to Misalign! Residual Self-Attentive Feature Alignment for Adapting Object Detectors

Vaishnavi Khindkar, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, Rohit Saluja, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3632-3642

Advancements in adaptive object detection can lead to tremendous improvements in applications like autonomous navigation, as they alleviate the distributional shifts along the detection pipeline. Prior works adopt adversarial learning to align image features at global and local levels, yet the instance-specific misalignment persists. Also, adaptive object detection remains challenging due to visual diversity in background scenes and intricate combinations of objects. Motivated by structural importance, we aim to attend prominent instance-specific regions, overcoming the feature misalignment issue. We propose a novel resIduaL seLf-attentive featUre alignMEnt (ILLUME) method for adaptive object detection. ILLUME comprises Self-Attention Feature Map (SAFM) module that enhances structural attention to object-related regions and thereby generates domain invariant features. Our approach significantly reduces the domain distance with the improved feature alignment of the instances. Qualitative results demonstrate the ability of ILLUME to attend important object instances required for alignment. Experimental results on several benchmark datasets show that our method outperforms the existing state-of-the-art approaches.

****

EZCrop: Energy-Zoned Channels for Robust Output Pruning

Rui Lin, Jie Ran, Dongpeng Wang, King Hung Chiu, Ngai Wong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 19-28

Recent results have revealed an interesting observation in a trained convolutional neural network (CNN), namely, the rank of a feature map channel matrix remains surprisingly constant despite the input images. This has led to an effective rank-based channel pruning algorithm, yet the constant rank phenomenon remains mysterious and unexplained. This work aims at demystifying and interpreting such rank behavior from a frequency-domain perspective, which as a bonus suggests an extremely efficient Fast Fourier Transform (FFT)-based metric for measuring channel importance without explicitly computing its rank. We achieve remarkable CNN channel pruning based on this analytically sound and computationally efficient metric, and adopt it for repetitive pruning to demonstrate robustness via our scheme named Energy-Zoned Channels for Robust Output Pruning (EZCrop), which shows consistently better results than other state-of-the-art channel pruning methods. The codes and Appendix are publicly available at: https://github.com/ruilin0212/EZCrop.

****

Geometrically Adaptive Dictionary Attack on Face Recognition

Junyoung Byun, Hyojun Go, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3021-3030

CNN-based face recognition models have brought remarkable performance improvement, but they are vulnerable to adversarial perturbations. Recent studies have sho wn that adversaries can fool the models even if they can only access the models' hard-label output. However, since many queries are needed to find imperceptible adversarial noise, reducing the number of queries is crucial for these attacks. In this paper, we point out two limitations of existing decision-based black-box attacks. We observe that they waste queries for background noise optimization, and they do not take advantage of adversarial perturbations generated for other images. We exploit 3D face alignment to overcome these limitations and propose a general strategy for query-efficient black-box attacks on face recognition named Geometrically Adaptive Dictionary Attack (GADA). Our core idea is to create an adversarial perturbation in the UV texture map and project it onto the face in the image. It greatly improves query efficiency by limiting the perturbation search space to the facial area and effectively recycling previous perturbations. We apply the GADA strategy to two existing attack methods and show overwhelming performance improvement in the experiments on the LFW and CPLFW datasets. Furthermore, we also present a novel attack strategy that can circumvent query similarity-based stateful detection that identifies the process of query-based black-box attacks.

****

A Riemannian Framework for Analysis of Human Body Surface

Emery Pierson, Mohamed Daoudi, Alice-Barbara Tumpach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2991-3000

We propose a novel framework for comparing 3D human shapes under the change of shape and pose. This problem is challenging since 3D human shapes vary significantly across subjects and body postures. We solve this problem by using a Riemannian approach. Our core contribution is the mapping of the human body surface to the space of metrics and normals. We equip this space with a family of Riemannian metrics, called Ebin (or DeWitt) metrics. We treat a human body surface as a point in a "shape space" equipped with a family of Riemmanian metrics. The family of metrics is invariant under rigid motions and reparametrizations; hence it induces a metric on the "shape space" of surfaces. Using the alignment of human bodies with a given template, we show that this family of metrics allows us to distinguish the changes in shape and pose. The proposed framework has several advantages. First, we define a family of metrics with desired invariant properties for the comparison of human shape. Second, we present an efficient framework to compute geodesic paths between human shape given the chosen metric. Third, this framework provides some basic tools for statistical shape analysis of human body surfaces. Finally, we demonstrate the utility of the proposed framework in pose and shape retrieval of human body.

****

Hyper-Convolution Networks for Biomedical Image Segmentation

Tianyu Ma, Adrian V. Dalca, Mert R. Sabuncu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1933-1942

The convolution operation is a central building block of neural network architectures widely used in computer vision. The size of the convolution kernels determines both the expressiveness of convolutional neural networks (CNN), as well as the number of learnable parameters. Increasing the network capacity to capture rich pixel relationships requires increasing the number of learnable parameters, often leading to overfitting and/or lack of robustness. In this paper, we propose a powerful novel building block, the hyper-convolution, which implicitly represents the convolution kernel as a function of kernel coordinates. Hyper-convolutions enable decoupling the kernel size, and hence its receptive field, from the number of learnable parameters. In our experiments, focused on challenging biomedical image segmentation tasks, we demonstrate that replacing regular convolutions with hyper-convolutions leads to more efficient architectures that achieve improved accuracy. Our analysis also shows that learned hyper-convolutions are naturally regularized, which can offer better generalization performance. We believe that hyper-convolutions can be a powerful building block in future neural network architectures solving computer vision tasks. We provide all of our code here: https://github.com/tym002/Hyper-Convolution

****

AuxAdapt: Stable and Efficient Test-Time Adaptation for Temporally Consistent Video Semantic Segmentation

Yizhe Zhang, Shubhankar Borse, Hong Cai, Fatih Porikli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2339-2348

In video segmentation, generating temporally consistent results across frames is as important as achieving frame-wise accuracy. Existing methods rely either on optical flow regularization or fine-tuning with test data to attain temporal consistency. However, optical flow is not always avail-able and reliable. Besides, it is expensive to compute. Fine-tuning the original model in test time is cost sensitive. This paper presents an efficient, intuitive, and unsupervised online adaptation method, AuxAdapt, for improving the temporal consistency of most neural network models. It does not require optical flow and only takes one pass of the video. Since inconsistency mainly arises from the model's uncertainty in its output, we propose an adaptation scheme where the model learns from its own segmentation decisions as it streams a video, which allows producing more confident and temporally consistent labeling for similarly-looking pixels across frames. For stability and efficiency, we leverage a small auxiliary segmentation network (AuxNet) to assist with this adaptation. More specifically, AuxNet readjusts the decision of the original segmentation network (Main-Net) by adding its own estimations to that of MainNet. At every frame, only AuxNet is updated via back-propagation while keeping MainNet fixed. We extensively evaluate our test-time adaptation approach on standard video benchmarks, including Cityscapes, CamVid, and KITTI. The results demonstrate that our approach provides label-wise accurate, temporally consistent, and computationally efficient adaptation (5+ folds overhead reduction comparing to state-of-the-art test-time adaptation methods).

****

Busy-Quiet Video Disentangling for Video Classification

Guoxi Huang, Adrian G. Bors; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1341-1350

In video data, busy motion details from moving regions are conveyed within a specific frequency bandwidth in the frequency domain. Meanwhile, the rest of the frequencies of video data are encoded with quiet information with substantial redundancy, which causes low processing efficiency in existing video models that take as input raw RGB frames. In this paper, we consider allocating intenser computation for the processing of the important busy information and less computation for that of the quiet information. We design a trainable Motion Band-Pass Module (MBPM) for separating busy information from quiet information in raw video data. By embedding the MBPM into a two-pathway CNN architecture, we define a Busy-Quiet Net (BQN). The efficiency of BQN is determined by avoiding redundancy in the feature space processed by the two pathways: one operating on Quiet features of low-resolution, while the other processes Busy features. The proposed BQN outperforms many recent video processing models on Something-Something V1, Kinetics400, UCF101 and HMDB51 datasets.

****