3D Change Localization and Captioning From Dynamic Scans of Indoor Scenes
Yue Qiu, Shintaro Yamamoto, Ryosuke Yamada, Ryota Suzuki, Hirokatsu Kataoka, Ken ji Iwata, Yutaka Satoh; Proceedings of the IEEE/CVF Winter Conference on Applica tions of Computer Vision (WACV), 2023, pp. 1176-1185
Daily indoor scenes often involve constant changes due to human activities. To r ecognize scene changes, existing change captioning methods focus on describing c hanges from two images of a scene. However, to accurately perceive and appropria tely evaluate physical changes and then identify the geometry of changed objects , recognizing and localizing changes in 3D space is crucial. Therefore, we propo se a task to explicitly localize changes in 3D bounding boxes from two point clo uds and describe detailed scene changes, including change types, object attribut es, and spatial locations. Moreover, we create a simulated dataset with various scenes, allowing generating data without labor costs. We further propose a frame work that allows different 3D object detectors to be incorporated in the change detection process, after which captions are generated based on the correlations of different change regions. The proposed framework achieves promising results i n both change detection and captioning. Furthermore, we also evaluated on data c ollected from real scenes. The experiments show that pretraining on the proposed dataset increases the change detection accuracy by +12.8% (mAP0.25) when applie d to real-world data. We believe that our proposed dataset and discussion could provide both a new benchmark and insights for future studies in scene change und erstanding.
**********

Panoptic-Aware Image-to-Image Translation
Liyun Zhang, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, Haruo Takemura; Proceedings of the IEEE/CVF Winter Conference on Appl ications of Computer Vision (WACV), 2023, pp. 259-268
Despite remarkable progress in image translation, the complex scene with multipl e discrepant objects remains a challenging problem. The translated images have l ow fidelity and tiny objects in fewer details causing unsatisfactory performance in object recognition. Without thorough object perception (i.e., bounding boxes , categories, and masks) of images as prior knowledge, the style transformation of each object will be difficult to track in translation. We propose panoptic-aw are generative adversarial networks (PanopticGAN) for image-to-image translation together with a compact panoptic segmentation dataset. The panoptic perception (i.e., foreground instances and background semantics of the image scene) is extr acted to achieve alignment between object content codes of the input domain and panoptic-level style codes sampled from the target style space, then refined by a proposed feature masking module for sharping object boundaries. The image-leve l combination between content and sampled style codes is also merged for higher fidelity image generation. Our proposed method was systematically compared with different competing methods and obtained significant improvement in both image q uality and object recognition performance.
**********

Partially Calibrated Semi-Generalized Pose From Hybrid Point Correspondences
Snehal Bhayani, Torsten Sattler, Viktor Larsson, Janne Heikkilä, Zuzana Kukelova ; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi on (WACV), 2023, pp. 2882-2891
In this paper we study the problem of estimating the semi-generalized pose of a partially calibrated camera, i.e., the pose of a perspective camera with unknown focal length w.r.t. a generalized camera, from a hybrid set of 2D-2D and 2D-3D point correspondences. We study all possible camera configurations within the ge neralized camera system. To derive practical solvers to previously unsolved chal lenging configurations, we test different parameterizations as well as different solving strategies based on the state-of-the-art methods for generating efficie nt polynomial solvers. We evaluate the three most promising solvers, i.e., the H 51f solver with five 2D-2D correspondences and one 2D-3D correspondence viewed b y the same camera inside generalized camera, the H32f solver with three 2D-2D an d two 2D-3D correspondences, and the H13f solver with one 2D-2D and three 2D-3D correspondences, on synthetic and real data. We show that in the presence of noi

se in the 3D points these solvers provide better estimates than the correspondin g absolute pose solvers.
**********

Elimination of Non-Novel Segments at Multi-Scale for Few-Shot Segmentation
Alper Kayaba■■, Gülin Tüfekci, ■lkay Ulusoy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2559-2567
Few-shot segmentation aims to devise a generalizing model that segments query im ages from unseen classes during training with the guidance of a few support imag es whose class tally with the class of the query. There exist two domain-specifi c problems mentioned in the previous works, namely spatial inconsistency and bia s towards seen classes. Taking the former problem into account, our method compa res the support feature map with the query feature map at multi scales to become scale-agnostic. As a solution to the latter problem, a supervised model, called as base learner, is trained on available classes to accurately identify pixels belonging to seen classes. Hence, subsequent meta learner has a chance to discar d areas belonging to seen classes with the help of an ensemble learning model th at coordinates meta learner with the base learner. We simultaneously address the se two vital problems for the first time and achieve state-of-the-art performanc es on both PASCAL-5i and COCO-20i datasets.
**********

Continual Learning With Dependency Preserving Hypernetworks
Dupati Srikar Chandra, Sakshi Varshney, P. K. Srijith, Sunil Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 202 3, pp. 2339-2348
Humans learn continually throughout their lifespan by accumulating diverse knowl edge and fine-tuning it for future tasks. When presented with a similar goal, ne ural networks suffer from catastrophic forgetting if data distributions across s equential tasks are not stationary over the course of learning. An effective app roach to address such continual learning (CL) problems is to use hypernetworks w hich generate task dependent weights for a target network. However, the continua l learning performance of existing hypernetwork based approaches are affected by the assumption of independence of the weights across the layers in order to mai ntain parameter efficiency. To address this limitation, we propose a novel appro ach that uses a dependency preserving hypernetwork to generate weights for the t arget network while also maintaining the parameter efficiency. We propose to use recurrent neural network (RNN) based hypernetwork that can generate layer weigh ts efficiently while allowing for dependencies across them. In addition, we prop ose novel regularisation and network growth techniques for the RNN based hyperne twork to further improve the continual learning performance. To demonstrate the effectiveness of the proposed methods, we conducted experiments on several image classification continual learning tasks and settings. We found that the propose d methods based on the RNN hypernetworks outperformed the baselines in all these CL settings and tasks.
**********

Learning How to MIMIC: Using Model Explanations To Guide Deep Learning Training
Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1461-1470
Healthcare is seen as one of the most influential applications of Deep Learning (DL). Increasingly, DL models are applied in healthcare settings with seemingly high levels of performance on-par with medical experts. Yet, very few are deploy ed into real-life scenarios with variable success rate. One of the main reasons for this is the lack of trust in those models by medical professionals driven by the black-box nature of the deployed models. Numerous explainable techniques ha ve been developed to alleviate this issue by providing a view on how the model r eached a given decision. Recent studies have shown that those explanations can e xpose the models' reliance on areas of the feature space that has no justifiable medical interpretation, widening the gap with the medical experts. In this pape r we evaluate the deviation of saliency maps produced by DL classification model s from radiologist's eye-gaze while they study the MIMIC-CXR-EGD images, and we

propose a novel model architecture that utilises model explanations during train ing only (i.e. not during inference) to improve the overall plausibility of the model explanations. We substantially improve the similarity between the model's explanations and radiologists' eye-gaze data, reducing Kullback-Leibler Divergen ce by 90% and increasing Normalised Scanpath Saliency by 216%. We argue that thi s significant improvement is an important step towards building more robust and interpretable DL solutions in healthcare.
**********

Learning by Hallucinating: Vision-Language Pre-Training With Weak Supervision
Tzu-Jui Julius Wang, Jorma Laaksonen, Tomas Langer, Heikki Arponen, Tom E. Bisho p; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vis ion (WACV), 2023, pp. 1073-1083
Weakly-supervised vision-language (V-L) pre-training (W-VLP) aims at learning cr oss-modal alignment with little or no paired data, such as aligned images and ca ptions. Recent W-VLP methods, which pair visual features with object tags, help achieve performances comparable with some VLP models trained with aligned pairs in various V-L downstream tasks. This, however, is not the case in cross- modal retrieval (XMR). We argue that the learning of such a W-VLP model is curbed and biased by the object tags of limited semantics. We address the lack of paired V- L data for model supervision with a novel Visual Vocabulary based Feature Halluc inator (WFH), which is trained via weak supervision as a W-VLP model, not requir ing images paired with captions. WFH generates visual hallucinations from texts, which are then paired with the originally unpaired texts, allowing more diverse interactions across modalities. Empirically, WFH consistently boosts the prior W-VLP works, e.g. U-VisualBERT (U-VB), over a variety of V-L tasks, i.e. XMR, Vi sual Question Answering, etc. Notably, benchmarked with recall@ 1,5,10 , it cons istently improves U-VB on image-to-text and text-to-image retrieval on two popul ar datasets Flickr30K and MSCOCO. Meanwhile, it gains by at least 14.5% in cross -dataset generalization tests on these XMR tasks. Moreover, in other V-L downstr eam tasks considered, our WFH models are on par with models trained with paired V-L data, revealing the utility of unpaired data. These results demonstrate grea ter generalization of the proposed W-VLP model with WFH.
**********

Self-Supervised Relative Pose With Homography Model-Fitting in the Loop
Bruce R. Muller, William A. P. Smith; Proceedings of the IEEE/CVF Winter Confere nce on Applications of Computer Vision (WACV), 2023, pp. 5705-5714
We propose a self-supervised method for relative pose estimation for road scenes . By exploiting the approximate planarity of the local ground plane, we can extr act a self-supervision signal via cross-projection between images using a homogr aphy derived from estimated ground-relative pose. We augment cross-projected per ceptual loss by including classical image alignment in the network training loop . We use pretrained semantic segmentation and optical flow to extract ground pla ne correspondences between approximately aligned images and RANSAC to find the b est fitting homography. By decomposing to ground-relative pose, we obtain pseudo labels that can be used for direct supervision. We show that this extremely sim ple geometric model is competitive for visual odometry with much more complex se lf-supervised methods that must learn depth estimation in conjunction with relat ive pose. Code and result videos: github.com/brucemuller/homographyVO.
**********

CAST: Conditional Attribute Subsampling Toolkit for Fine-Grained Evaluation
Wes Robbins, Steven Zhou, Aman Bhatta, Chad Mello, Vítor Albiero, Kevin W. Bowye r, Terrance E. Boult; Proceedings of the IEEE/CVF Winter Conference on Applicati ons of Computer Vision (WACV), 2023, pp. 919-929
Thorough evaluation is critical for developing models that are fair and robust. In this work, we describe the Conditional Attribute Subsampling Toolkit (CAST) f or selecting data subsets for fine-grained scientific evaluations. Our toolkit e fficiently filters data given an arbitrary number of conditions for metadata att ributes. The purpose of the toolkit is to allow researchers to easily to evaluat e models on targeted test distributions. The functionality of CAST is demonstrat ed on the WebFace42M face Recognition dataset. We calculate over 50 attributes f

or this dataset including race, image quality, facial features, and accessories. Using our toolkit, we create over a hundred test sets conditioned on one or multiple attributes. Results are presented for subsets of various demographics and image quality ranges. Using eleven different subsets, we build a face recognition 1:1 verification benchmark called C11 that exclusively contains pairs that are near the decision threshold. Evaluation on C11 with state-of-the-art methods demonstrates the suitability of the proposed benchmark. The toolkit is publicly available at https://github.com/WesRobbins/CAST.

**********

Seq-UPS: Sequential Uncertainty-Aware Pseudo-Label Selection for Semi-Supervised Text Recognition

Gaurav Patel, Jan P. Allebach, Qiang Qiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6180-6190

This paper looks at semi-supervised learning (SSL) for image-based text recognition. One of the most popular SSL approaches is pseudo-labeling (PL). PL approaches assign labels to unlabeled data before re-training the model with a combination of labeled and pseudo-labeled data. However, PL methods are severely degraded by noise and are prone to over-fitting to noisy labels, due to the inclusion of erroneous high confidence pseudo-labels generated from poorly calibrated models, thus, rendering threshold-based selection ineffective. Moreover, the combinatorial complexity of the hypothesis space and the error accumulation due to multiple incorrect autoregressive steps posit pseudo-labeling challenging for sequential self-training. To this end, we propose a pseudo-label generation and an uncertainty-based data selection framework for semi-supervised text recognition. We first use Beam-Search inference to yield highly probable hypotheses to assign pseudo-labels to the unlabelled examples. Then we adopt an ensemble of models, sampled by applying dropout, to obtain a robust estimate of the uncertainty associated with the prediction, considering both the character-level and word-level predictive distribution to select good quality pseudo-labels. Extensive experiments on several benchmark handwriting and scene-text datasets show that our method outperforms the baseline approaches and the previous state-of-the-art semi-supervised text-recognition methods.

**********

Text and Image Guided 3D Avatar Generation and Manipulation

Zehranaz Canfes, M. Furkan Atasoy, Alara Dirik, Pinar Yanardag; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4421-4431

The manipulation of latent space has recently become an interesting topic in the field of generative models. Recent research shows that latent directions can be used to manipulate images towards certain attributes. However, controlling the generation process of 3D generative models remains a challenge. In this work, we propose a novel 3D manipulation method that can manipulate both the shape and texture of the model using text or image-based prompts such as 'a young face' or 'a surprised face'. We leverage the power of Contrastive Language-Image Pre-training (CLIP) model and a pre-trained 3D GAN model designed to generate face avatars, and create a fully differentiable rendering pipeline to manipulate meshes. More specifically, our method takes an input latent code and modifies it such that the target attribute specified by a text or image prompt is present or enhanced, while leaving other attributes largely unaffected. Our method requires only 5 minutes per manipulation, and we demonstrate the effectiveness of our approach with extensive results and comparisons.

**********

RAST: Restorable Arbitrary Style Transfer via Multi-Restoration

Yingnan Ma, Chenqiu Zhao, Xudong Li, Anup Basu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 331-340

Arbitrary style transfer aims at reproducing the target image with provided artistic or photo-realistic styles. Even though existing approaches can successfully transfer style information, arbitrary style transfer still faces many challenges, such as the content leak issue. To be specific, the embedding of artistic style can lead to content changes. In this paper, we solve the content leak problem

from the perspective of image restoration. In particular, an iterative architecture is proposed to achieve the restorable arbitrary style transfer (RAST), which can realize the transmission of both content and style information through the multi-restorations. We control the content-style balance in stylized images by the accuracy of image restoration. In order to ensure the effectiveness of the proposed RAST architecture, we design two novel loss functions: multi-restoration loss and style difference loss. In addition, we propose a new quantitative evaluation method to measure content preservation performance and style embedding performance. Comprehensive experiments comparing with state-of-the-art methods demonstrate that our proposed architecture can produce stylized images with superior performance on content preservation and style embedding.
**********

Surface Normal Estimation From Optimized and Distributed Light Sources Using DNN-Based Photometric Stereo
Takafumi Iwaguchi, Hiroshi Kawasaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 311-320
Photometric stereo (PS) is a major technique to recover surface normal for each pixel. However, since it assumes Lambertian surface and directional light to estimate the value, a large number of images are usually required to avoid the effects of outliers and noise. In this paper, we propose a technique to reduce the number of images by using distributed light sources, where the patterns are optimized by a deep neural network (DNN). In addition, to efficiently realize the distributed light, we use an optical diffuser with a video projector, where the diffuser is illuminated by the projector from behind, the illuminated area on the diffuser works as if an arbitrary-shaped area light. To estimate the surface normal using the distributed light source, we propose a near-light photometric stereo (NLPS) using DNN. Since optimization of the pattern of distributed light is achieved by a differentiable renderer, it is connected with NLPS network, achieving end-to-end learning. The experiments are conducted to show the successful estimation of the surface normal by our method from a small number of images.
**********

UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval
Andreas Specker, Mickael Cormier, Jürgen Beyerer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 981-990
Recognizing soft-biometric pedestrian attributes is essential in video surveillance and fashion retrieval. Recent works show promising results on single datasets. Nevertheless, the generalization ability of these methods under different attribute distributions, viewpoints, varying illumination, and low resolutions remains rarely understood due to strong biases and varying attributes in current datasets. To close this gap and support a systematic investigation, we present UPAR, the Unified Person Attribute Recognition Dataset. It is based on four well-known person attribute recognition datasets: PA100K, PETA, RAPv2, and Market1501. We unify those datasets by providing 3,3M additional annotations to harmonize 40 important binary attributes over 12 attribute categories across the datasets. We thus enable research on generalizable pedestrian attribute recognition as well as attribute-based person retrieval for the first time. Due to the vast variance of the image distribution, pedestrian pose, scale, and occlusion, existing approaches are greatly challenged both in terms of accuracy and efficiency. Furthermore, we develop a strong baseline for PAR and attribute-based person retrieval based on a thorough analysis of regularization methods. Our models achieve state-of-the-art performance in cross-domain and specialization settings on PA100k, PETA, RAPv2, Market1501-Attributes, and UPAR. We believe UPAR and our strong baseline will contribute to the artificial intelligence community and promote research on large-scale, generalizable attribute recognition systems.
**********

Instance-Dependent Noisy Label Learning via Graphical Modelling
Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, Gustavo Carneiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2288-2298
Noisy labels are unavoidable yet troublesome in the ecosystem of deep learning b

ecause models can easily overfit them. There are many types of label noise, such as symmetric, asymmetric and instance-dependent noise (IDN), with IDN being the only type that depends on image information. Such dependence on image information makes IDN a critical type of label noise to study, given that labelling mistakes are caused in large part by insufficient or ambiguous information about the visual classes present in images. Aiming to provide an effective technique to address IDN, we present a new graphical modelling approach called InstanceGM, that combines discriminative and generative models. The main contributions of InstanceGM are: i) the use of the continuous Bernoulli distribution to train the generative model, offering significant training advantages, and ii) the exploration of a state-of-the-art noisy-label discriminative classifier to generate clean labels from instance-dependent noisy-label samples. InstanceGM is competitive with current noisy-label learning approaches, particularly in instance-dependent noise benchmarks using synthetic and real-world datasets, where our method shows better accuracy than the competitors in most experiments.

**********

On the Importance of Denoising When Learning To Compress Images

Benoit Brummer, Christophe De Vleeschouwer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2440-2448

Image noise is ubiquitous in photography. However, image noise is not compressible nor desirable, thus attempting to convey the noise in compressed image bitstreams yields sub-par results in both rate and distortion. We propose to explicitly learn the image denoising task when training the codec. Therefore, we leverage the Natural Image Noise Dataset, which offers a wide variety of scenes captured with various noise levels. Given this training set, we show that a single model trained based on a mixture of images with variable noise levels appears to yield best-in-class results with both noisy and clean images, achieving better rate-distortion than a compression-only model or even than a pair of denoising-then-compression models with almost one order of magnitude fewer GMac operations.

**********

AdaNorm: Adaptive Gradient Norm Correction Based Optimizer for CNNs

Shiv Ram Dubey, Satish Kumar Singh, Bidyut Baran Chaudhuri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5284-5293

The stochastic gradient descent (SGD) optimizers are generally used to train the convolutional neural networks (CNNs). In recent years, several adaptive momentum based SGD optimizers have been introduced, such as Adam, diffGrad, Radam and AdaBelief. However, the existing SGD optimizers do not exploit the gradient norm of past iterations and lead to poor convergence and performance. In this paper, we propose a novel AdaNorm based SGD optimizers by correcting the norm of gradient in each iteration based on the adaptive training history of gradient norm. By doing so, the proposed optimizers are able to maintain high and representive gradient throughout the training and solves the low and atypical gradient problems. The proposed concept is generic and can be used with any existing SGD optimizer. We show the efficacy of the proposed AdaNorm with four state-of-the-art optimizers, including Adam, diffGrad, Radam and AdaBelief. We depict the performance improvement due to the proposed optimizers using three CNN models, including VGG16, ResNet18 and ResNet50, on three benchmark object recognition datasets, including CIFAR10, CIFAR100 and TinyImageNet.

**********

Aerial Image Dehazing With Attentive Deformable Transformers

Ashutosh Kulkarni, Subrahmanyam Murala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6305-6314

Aerial imagery is widely utilized in visual data dependent applications such as military surveillance, earthquake assessment, etc. For these applications, minute texture in the aerial image are essential as any disturbance can cause inaccurate prediction. However, atmospheric haze severely reduces the visibility of the scene to be analysed, and hence takes a toll on accuracy of higher level applications. Existing methods either utilize additional prior while training, or produce sub-optimal outputs on different densities of haze degradation, due to absen

ce of local and global dependencies in the extracted features. Therefore, it is essential to have a texture preserving algorithm for aerial image dehazing. In light of this, we propose a work that introduces a novel deformable multi-head attention with spatially attentive offset extraction based solution for aerial image dehazing. Here, the deformable multi-head attention is introduced to reconstruct fine level texture in the restored image. We also introduce spatially attentive offset extractor in the deformable convolution for focusing on relevant contextual information. Further, edge boosting skip connections are proposed for effectively passing edge features from shallow layers to deeper layers of the network. Thorough experimentation on synthetic as well as real-world data, along with extensive ablation study, demonstrate that the proposed method outperforms the prevailing works on aerial image dehazing. The code is provided at https://github.com/ AshutoshKulkarni4998/AIDTransformer.

**********

Evaluating Generative Networks Using Gaussian Mixtures of Image Features

Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G. Baraniuk, Michael J. Henry; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 279-288

We develop a measure for evaluating the performance of generative networks given two sets of images. A popular performance measure currently used to do this is the Frechet Inception Distance (FID). FID assumes that images featurized using the penultimate layer of Inception-v3 follow a Gaussian distribution, an assumption which cannot be violated if we wish to use FID as a metric. However, we show that Inception-v3 features of the ImageNet dataset are not Gaussian; in particular, every single marginal is not Gaussian. To remedy this problem, we model the featurized images using Gaussian mixture models (GMMs) and compute the 2-Wasserstein distance restricted to GMMs. We define a performance measure, which we call WaM, on two sets of images by using Inception-v3 (or another classifier) to featurize the images, estimate two GMMs, and use the restricted 2-Wasserstein distance to compare the GMMs. We experimentally show the advantages of WaM over FID, including how FID is more sensitive than WaM to imperceptible image perturbations. By modelling the non-Gaussian features obtained from Inception-v3 as GMMs and using a GMM metric, we can more accurately evaluate generative network performance.

**********

Sparsity Agnostic Depth Completion

Andrea Conti, Matteo Poggi, Stefano Mattoccia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5871-5880

We present a novel depth completion approach agnostic to the sparsity of depth points, that is very likely to vary in many practical applications. State-of-the-art approaches yield accurate results only when processing a specific density and distribution of input points, i.e. the one observed during training, narrowing their deployment in real use cases. On the contrary, our solution is robust to uneven distributions and extremely low densities never witnessed during training. Experimental results on standard indoor and outdoor benchmarks highlight the robustness of our framework, achieving accuracy comparable to state-of-the-art methods when tested with density and distribution equal to the training one while being much more accurate in the other cases. Our pretrained models and further material are available in our project page.

**********

MovieCLIP: Visual Scene Recognition in Movies

Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, Shrikanth Narayanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2083-2092

Longform media such as movies have complex narrative structures, with events spanning a rich variety of ambient visual scenes. Domain-specific challenges associated with visual scenes in movies include transitions, person coverage, and a wide array of real-life and fictional scenarios. Existing visual scene datasets in movies have limited taxonomies and don't consider the visual scene transition w

ithin movie clips. In this work, we address the problem of visual scene recognit ion in movies by first automatically curating a new and extensive movie-centric taxonomy of 179 scene labels derived from movie scripts and auxiliary web-based video datasets. Instead of manual annotations which can be expensive, we use CLI P to weakly label 1.12 million shots from 32K movie clips based on our proposed taxonomy. We provide baseline visual models trained on the weakly labeled datase t called MovieCLIP and evaluate them on an independent dataset verified by human raters. We show that leveraging features from models pretrained on MovieCLIP be nefits downstream tasks such as multi-label scene and genre classification of we b videos and movie trailers.

**********

Dynamic Re-Weighting for Long-Tailed Semi-Supervised Learning

Hanyu Peng, Weiguo Pian, Mingming Sun, Ping Li; Proceedings of the IEEE/CVF Wint er Conference on Applications of Computer Vision (WACV), 2023, pp. 6464-6474

The high demand for labeled data that characterizes deep learning is very labor- intensive. Semi-supervised Learning (SSL), acting as one of the breakthroughs, a llows for the avoidance of this labeling loss thanks to its small amount of labe led data, alongside extracting information from a large amount of unlabeled data . And there is hope that the same performance for SSL can be achieved when compa red to supervised learning methods. Regrettably, the research community has ofte n developed SSL regarding the nature of a balanced data set; in contrast, real d ata is often imbalanced or even long-tailed. The need to study SSL under imbalan ce is therefore critical. In this paper, we shall essentially extend FixMatch (a SSL method) to the imbalanced case. We find that the unlabeled data is as well highly imbalanced during the training process; in this respect we propose a re-w eighting solution based on the effective number. Furthermore, since prediction u ncertainty leads to temporal variations in the number of pseudo-labels, we are i nnovative in proposing a dynamic re-weighting scheme on the unlabeled data. The simplicity and validity of our method are backed up by strong experimental evide nce. Especially on CIFAR-10, CIFAR-100, ImageNet127 data sets, our approach prov ides the strongest results against previous methods across various scales of imb alance.

**********

Grounding Scene Graphs on Natural Images via Visio-Lingual Message Passing

Aditay Tripathi, Anand Mishra, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4391-4400

This paper presents a framework for jointly grounding objects that follow certai n semantic relationship constraints given in a scene graph. A typical natural sc ene contains several objects, often exhibiting visual relationships of varied co mplexities between them. These inter-object relationships provide strong context ual cues towards improving grounding performance compared to a traditional objec t query-only-based localization task. A scene graph is an efficient and structur ed way to represent all the objects and their semantic relationships in the imag e. In an attempt towards bridging these two modalities representing scenes and u tilizing contextual information for improving object localization, we rigorously study the problem of grounding scene graphs on natural images. To this end, we propose a novel graph neural network-based approach referred to as Visio-Lingual Message PAssing Graph Neural Network (VL-MPAG Net). In VL-MPAG Net, we first co nstruct a directed graph with object proposals as nodes and an edge between a pa ir of nodes representing a plausible relation between them. Then a three-step in ter-graph and intra-graph message passing is performed to learn the context-depe ndent representation of the proposals and query objects. These object representa tions are used to score the proposals to generate object localization. The propo sed method significantly outperforms the baselines on four public datasets.

**********

Improving Multi-Fidelity Optimization With a Recurring Learning Rate for Hyperpa rameter Tuning

HyunJae Lee, Gihyeon Lee, Junhwan Kim, Sungjun Cho, Dohyun Kim, Donggeun Yoo; Pr oceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision ( WACV), 2023, pp. 2309-2318

Despite the evolution of Convolutional Neural Networks (CNNs), their performance is surprisingly dependent on the choice of hyperparameters. However, it remains challenging to efficiently explore large hyperparameter search space due to the long training times of modern CNNs. Multi-fidelity optimization enables the exploration of more hyperparameter configurations given budget by early termination of unpromising configurations. However, it often results in selecting a sub-optimal configuration as training with the high-performing configuration typically converges slowly in an early phase. In this paper, we propose Multi-fidelity Optimization with a Recurring Learning rate (MORL) which incorporates CNNs' optimization process into multi-fidelity optimization. MORL alleviates the problem of slow-starter and achieves a more precise low-fidelity approximation. Our comprehensive experiments on general image classification, transfer learning, and semi-supervised learning demonstrate the effectiveness of MORL over other multi-fidelity optimization methods such as Successive Halving Algorithm (SHA) and Hyperband. Furthermore, it achieves significant performance improvements over hand-tuned hyperparameter configuration within a practical budget.

**********

TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection

Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, Xin Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4691-4700

In this paper, we propose a Temporal Identity Inconsistency Network (TI2Net), a Deepfake detector that focuses on temporal identity inconsistency. Specifically, TI2Net recognizes fake videos by capturing the dissimilarities of human faces among video frames of the same identity. Therefore, TI2Net is a reference-agnostic detector and can be used on unseen datasets. For a video clip of a given identity, identity information in all frames will first be encoded to identity vectors. TI2Net learns the temporal identity embedding from the temporal difference of the identity vectors. The temporal embedding, representing the identity inconsistency in the video clip, is finally used to determine the authenticity of the video clip. During training, TI2Net incorporates triplet loss to learn more discriminative temporal embeddings. We conduct comprehensive experiments to evaluate the performance of the proposed TI2Net. Experimental results indicate that TI2Net generalizes well to unseen manipulations and datasets with unseen identities. Besides, TI2Net also shows robust performance against compression and additive noise.

**********

ScoreNet: Learning Non-Uniform Attention and Augmentation for Transformer-Based Histopathological Image Classification

Thomas Stegmüller, Behzad Bozorgtabar, Antoine Spahr, Jean-Philippe Thiran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6170-6179

Progress in digital pathology is hindered by high-resolution images and the prohibitive cost of exhaustive localized annotations. The commonly used paradigm to categorize pathology images is patch-based processing, which often incorporates multiple instance learning MIL to aggregate local patch-level representations yielding image-level prediction. Nonetheless, diagnostically relevant regions may only take a small fraction of the whole tissue, and current MIL-based approaches often process images uniformly, discarding the inter-patches interactions. To alleviate these issues, we propose ScoreNet, a new efficient transformer that exploits a differentiable recommendation stage to extract discriminative image regions and dedicate computational resources accordingly. The proposed transformer leverages the local and global attention of a few dynamically recommended high-resolution regions at an efficient computational cost. We further introduce a novel mixing data-augmentation, namely ScoreMix, by leveraging the image's semantic distribution to guide the data mixing and produce coherent sample-label pairs. ScoreMix is embarrassingly simple and mitigates the pitfalls of previous augmentations, which assume a uniform semantic distribution and risk mislabeling the samples. Thorough experiments and ablation studies on three breast cancer histology datasets of Haematoxylin & Eosin (H&E) have validated the superiority of our ap

proach over prior arts, including transformer-based models on tumour regions-of-interest TRoIs classification. ScoreNet equipped with proposed ScoreMix augmentation demonstrates better generalization capabilities and achieves new state-of-the-art (SOTA) results with only 50% of the data compared to other mixing augmentation variants. Finally, ScoreNet yields high efficacy and outperforms SOTA efficient transformers, namely TransPath and SwinTransformer, with throughput around 3x and 4x higher than the aforementioned architectures, respectively.
**********

Cross-View Image Sequence Geo-Localization
Xiaohan Zhang, Waqas Sultani, Safwan Wshah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2914-2923
Cross-view geo-localization aims to estimate the GPS location of a query ground-view image by matching it to images from a reference database of geo-tagged aerial images. To address this challenging problem, recent approaches use panoramic ground-view images to increase the range of visibility. Although appealing, panoramic images are not readily available compared to the videos of limited Field-Of-View (FOV) images. In this paper, we present the first cross-view geo-localization method that works on a sequence of limited FOV images. Our model is trained end-to-end to capture the temporal structure that lies within the frames using the attention-based temporal feature aggregation module. To robustly tackle different sequences length and GPS noises during inference, we propose to use a sequential dropout scheme to simulate variant length sequences. To evaluate the proposed approach in realistic settings, we present a new large-scale dataset containing ground-view sequences along with the corresponding aerial-view images. Extensive experiments and comparisons demonstrate the superiority of the proposed approach compared to several competitive baselines.
**********

CoKe: Contrastive Learning for Robust Keypoint Detection
Yutong Bai, Angtian Wang, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 65-74
In this paper, we introduce a contrastive learning framework for keypoint detection (CoKe). Keypoint detection differs from other visual tasks where contrastive learning has been applied because the input is a set of images in which multiple keypoints are annotated. This requires the contrastive learning to be extended such that the keypoints are represented and detected independently, which enables the contrastive loss to make the keypoint features different from each other and from the background. Our approach has two benefits: It enables us to exploit the power of contrastive learning for keypoint detection, and by detecting each keypoint independently the detection becomes more robust to occlusion compared to holistic methods, such as stacked hourglass networks, which attempt to detect all keypoints jointly. Our CoKe framework introduces several technical innovations. In particular, we introduce: (i) A clutter bank to represent non-keypoint features; (ii) a keypoint bank that stores prototypical representations of keypoints to approximate the contrastive loss between keypoints; and (iii) a cumulative moving average update to learn the keypoint prototypes while training the feature extractor. Our experiments on a range of diverse datasets (PASCAL3D+, MPII, ObjectNet3D) show that our approach works as well, or better than, alternative methods for keypoint detection, even for human keypoints, for which the literature is vast. Moreover, we observe that CoKe is exceptionally robust to partial occlusion and previously unseen object poses.
**********

BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video
Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, Deva Ramanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1674-1683
Multiple existing benchmarks involve tracking and segmenting objects in video e.g., Video Object Segmentation (VOS) and Multi-Object Tracking and Segmentation (MOTS), but there is little interaction between them due to the use of disparate

benchmark datasets and metrics (e.g. \JnF, mAP, sMOTSA). As a result, published works usually target a particular benchmark, and are not easily comparable to ea ch another. We believe that the development of generalized methods that can tack le multiple tasks requires greater cohesion among these research sub-communities . In this paper, we aim to facilitate this by proposing BURST, a dataset which c ontains thousands of diverse videos with high-quality object masks, and an assoc iated benchmark with six tasks involving object tracking and segmentation in vid eo. All tasks are evaluated using the same data and comparable metrics, which en ables researchers to consider them in unison, and hence, more effectively pool k nowledge from different methods across different tasks. Additionally, we demonst rate several baselines for all tasks and show that approaches for one task can b e applied to another with a quantifiable and explainable performance difference. Dataset annotations are available at: https://github.com/Ali2500/BURST-benchmar k.

**********

Collaborative Multi-Teacher Knowledge Distillation for Learning Low Bit-Width De ep Neural Networks

Cuong Pham, Tuan Hoang, Thanh-Toan Do; Proceedings of the IEEE/CVF Winter Confer ence on Applications of Computer Vision (WACV), 2023, pp. 6435-6443

Knowledge distillation which learns a lightweight student model by distilling kn owledge from a cumbersome teacher model is an attractive approach for learning c ompact deep neural networks (DNNs). Recent works further improve student network performance by leveraging multiple teacher networks. However, most of the exist ing knowledge distillation-based multi-teacher methods use separately pretrained teachers. This limits the collaborative learning between teachers and the mutua l learning between teachers and student. Network quantization is another at- tra ctive approach for learning compact DNNs. However, most existing network quantiz ation methods are developed and evaluated without considering multi-teacher supp ort to enhance the performance of quantized student model. In this paper, we pro pose a novel framework that leverages both multi-teacher knowledge distillation and network quantization for learning low bit-width DNNs. The proposed method en courages both collaborative learning between quantized teachers and mutual learn ing between quantized teachers and quantized student. During learning process, a t corresponding layers, knowledge from teachers will form an importance-aware sh ared knowledge which will be used as input for teachers at subsequent layers and also be used to guide student. Our experimental results on CIFAR100 and ImageNe t datasets show that the compact quantized student models trained with our metho d achieve competitive results compared to other state-of-the-art methods, and in some cases, indeed surpass the full precision models.

**********

Knowing What To Label for Few Shot Microscopy Image Cell Segmentation

Youssef Dawoud, Arij Bouazizi, Katharina Ernst, Gustavo Carneiro, Vasileios Bela giannis; Proceedings of the IEEE/CVF Winter Conference on Applications of Comput er Vision (WACV), 2023, pp. 3568-3577

In microscopy image cell segmentation, it is common to train a deep neural netwo rk on source data, containing different types of microscopy images, and then fin e-tune it using a support set comprising a few randomly selected and annotated t raining target images. In this paper, we argue that the random selection of unla belled training target images to be annotated and included in the support set ma y not enable an effective fine-tuning process, so we propose a new approach to o ptimise this image selection process. Our approach involves a new scoring functi on to find informative unlabelled target images. In particular, we propose to me asure the consistency in the model predictions on target images against specific data augmentations. However, we observe that the model trained with source data sets does not reliably evaluate consistency on target images. To alleviate this problem, we propose novel self-supervised pretext tasks to compute the scores of unlabelled target images. Finally, the top few images with the least consistenc y scores are added to the support set for oracle (i.e., expert) annotation and l ater used to fine-tune the model to the target images. In our evaluations that i nvolve the segmentation of five different types of cell images, we demonstrate p

romising results on several target test sets compared to the random selection ap
proach as well as other selection approaches, such as Shannon's entropy and Mont
e-Carlo dropout. Our code will be made publicly available.
**********

SSFE-Net: Self-Supervised Feature Enhancement for Ultra-Fine-Grained Few-Shot Cl
ass Incremental Learning
Zicheng Pan, Xiaohan Yu, Miaohua Zhang, Yongsheng Gao; Proceedings of the IEEE/C
VF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6275-6
284
Ultra-Fine-Grained Visual Categorization (ultra-FGVC) has become a popular probl
em due to its great real-world potential for classifying the same or closely rel
ated species with very similar layouts. However, there present many challenges f
or the existing ultra-FGVC methods, firstly there are always not enough samples
in the existing ultra-FGVC datasets based on which the models can easily get ove
rfitting. Secondly, in practice, we are likely to find new species that we have
not seen before and need to add them to existing models, which is known as incre
mental learning. The existing methods solve these problems by Few-Shot Class Inc
remental Learning (FSCIL), but the main challenge of the FSCIL models on ultra-F
GVC tasks lies in their inferior discrimination detection ability since they usu
ally use low-capacity networks to extract features, which leads to insufficient
discriminative details extraction from ultra-fine-grained images. In this paper,
a self-supervised feature enhancement for the few-shot incremental learning net
work (SSFE-Net) is proposed to solve this problem. Specifically, a self-supervis
ed learning (SSL) and knowledge distillation (KD) framework is developed to enha
nce the feature extraction of the low-capacity backbone network for ultra-FGVC f
ew-shot class incremental learning tasks. Besides, we for the first time create
a series of benchmarks for FSCIL tasks on two public ultra-FGVC datasets and thr
ee normal fine-grained datasets, which will facilitate the development of the Ul
tra-FGVC community. Extensive experimental results on public ultra-FGVC datasets
and other state-of-the-art benchmarks consistently demonstrate the effectivenes
s of the proposed method.
**********

MEVID: Multi-View Extended Videos With Identities for Video Person Re-Identifica
tion
Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic
Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, Brian Clipp
; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi
on (WACV), 2023, pp. 1634-1643
In this paper, we present the Multi-view Extended Videos with Identities (MEVID)
dataset for large-scale, video person re-identification (ReID) in the wild. To
our knowledge, MEVID represents the most-varied video person ReID dataset, spann
ing an extensive indoor and outdoor environment across nine unique dates in a 73
-day window, various camera viewpoints, and entity clothing changes. Specificall
y, we label the identities of 158 unique people wearing 598 outfits taken from 8
,092 tracklets, average length of about 590 frames, seen in 33 camera views from
the very-large-scale MEVA person activities dataset. While other datasets have
more unique identities, MEVID emphasizes a richer set of information about each
individual, such as: 4 outfits/identity vs. 2 outfits/identity in CCVID, 33 view
points across 17 locations vs. 6 in 5 simulated locations for MTA, and 10 millio
n frames vs. 3 million for LS-VID. Being based on the MEVA video dataset, we als
o inherit data that is intentionally demographically balanced to the continental
United States. To accelerate the annotation process, we developed a semi-automa
tic annotation framework and GUI that combines state-of-the-art real-time models
for object detection, pose estimation, person ReID, and multi-object tracking.
We evaluate several state-of-the-art methods on MEVID challenge problems and com
prehensively quantify their robustness in terms of changes of outfit, scale, and
background location. Our quantitative analysis on the realistic, unique aspects
of MEVID shows that there are significant remaining challenges in video person
ReID and indicates important directions for future research.
**********

Audio-Visual Efficient Conformer for Robust Speech Recognition
Maxime Burchi, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2258-2267
End-to-end Automatic Speech Recognition (ASR) systems based on neural networks have seen large improvements in recent years. The availability of large scale hand-labeled datasets and sufficient computing resources made it possible to train powerful deep neural networks, reaching very low Word Error Rate (WER) on academic benchmarks. However, despite impressive performance on clean audio samples, a drop of performance is often observed on noisy speech. In this work, we propose to improve the noise robustness of the recently proposed Efficient Conformer Connectionist Temporal Classification (CTC)-based architecture by processing both audio and visual modalities. We improve previous lip reading methods using an Efficient Conformer back-end on top of a ResNet-18 visual front-end and by adding intermediate CTC losses between blocks. We condition intermediate block features on early predictions using Inter CTC residual modules to relax the conditional independence assumption of CTC-based models. We also replace the Efficient Conformer grouped attention by a more efficient and simpler attention mechanism that we call patch attention. We experiment with publicly available Lip Reading Sentences 2 (LRS2) and Lip Reading Sentences 3 (LRS3) datasets. Our experiments show that using audio and visual modalities allows to better recognize speech in the presence of environmental noise and significantly accelerate training, reaching lower WER with 4 times less training steps. Our Audio-Visual Efficient Conformer (AVEC) model achieves state-of-the-art performance, reaching WER of 2.3% and 1.8% on LRS2 and LRS3 test sets. Code and pretrained models are available at https://github.com/burchim/AVEC.
**********
DigiFace-1M: 1 Million Digital Face Images for Face Recognition
Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, Jingjing Shen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3526-3535
State-of-the-art face recognition models show impressive accuracy, achieving over 99.8% on Labeled Faces in the Wild (LFW) dataset. Such models are trained on large-scale datasets that contain millions of real human face images collected from the internet. Web-crawled face images are severely biased (in terms of race, lighting, make-up, etc) and often contain label noise. More importantly, the face images are collected without explicit consent, raising ethical concerns. To avoid such problems, we introduce a large-scale synthetic dataset for face recognition, obtained by rendering digital faces using a computer graphics pipeline. We first demonstrate that aggressive data augmentation can significantly reduce the synthetic-to-real domain gap. Having full control over the rendering pipeline, we also study how each attribute (e.g., variation in facial pose, accessories and textures) affects the accuracy. Compared to SynFace, a recent method trained on GAN-generated synthetic faces, we reduce the error rate on LFW by 52.5% (accuracy from 91.93% to 96.17%). By fine-tuning the network on a smaller number of real face images that could reasonably be obtained with consent, we achieve accuracy that is comparable to the methods trained on millions of real face images.
**********
Couplformer: Rethinking Vision Transformer With Coupling Attention
Hai Lan, Xihao Wang, Hao Shen, Peidong Liang, Xian Wei; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6475-6484
With the development of the self-attention mechanism, the Transformer model has demonstrated its outstanding performance in the computer vision domain. However, the massive computation brought from the full attention mechanism became a heavy burden for memory consumption. Sequentially, the limitation of memory consumption hinders the deployment of the Transformer model on the embedded system where the computing resources are limited. To remedy this problem, we propose a novel memory economy attention mechanism named Couplformer, which decouples the attention map into two sub-matrices and generates the alignment scores from spatial information. Our method enables the Transformer model to improve time and memory

efficiency while maintaining expressive power. A series of different scale image classification tasks are applied to evaluate the effectiveness of our model. The result of experiments shows that on the ImageNet-1K classification task, the Couplformer can significantly decrease 42% memory consumption compared with the regular Transformer. Meanwhile, it accesses sufficient accuracy requirements, which outperforms 0.56% on Top-1 accuracy and occupies the same memory footprint. Besides, the Couplformer achieves state-of-art performance in MS COCO 2017 object detection and instance segmentation tasks. As a result, the Couplformer can serve as an efficient backbone in visual tasks and provide a novel perspective on deploying attention mechanisms for researchers.

**********

Synthetic Latent Fingerprint Generator

André Brasil Vieira Wyzykowski, Anil K. Jain; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 971-980

Given a full fingerprint image (rolled or slap), we present CycleGAN models to generate multiple latent impressions of the same identity as the full print. Our models can control the degree of distortion, noise, blurriness and occlusion in the generated latent print images to obtain Good, Bad and Ugly latent image categories as introduced in the NIST SD27 latent database. The contributions of our work are twofold: (i) demonstrate the similarity of synthetically generated latent fingerprint images to crime scene latents in NIST SD27 and MSP databases as evaluated by the NIST NFIQ 2 quality measure and recognition accuracies obtained by a SOTA fingerprint matcher, and (ii) use of synthetic latents to augment small-size latent training databases in the public domain to improve the performance of DeepPrint, a SOTA fingerprint matcher designed for rolled to rolled fingerprint matching on three latent databases (NIST SD27, NIST SD302, and IIITD-SLF). As an example, with synthetic latent data augmentation, the Rank-1 retrieval performance of DeepPrint is improved from 15.50% to 29.07% on challenging NIST SD27 latent database. Our approach for generating synthetic latent fingerprints can be used to improve the recognition performance of any latent matcher and its individual components (e.g., enhancement, segmentation and feature extraction). https://prip-lab.github.io/Synthetic-Latent-Fingerprint-Generator/

**********

Accumulated Trivial Attention Matters in Vision Transformers on Small Datasets

Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, Guanghui Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3984-3992

Vision Transformers has demonstrated competitive performance on computer vision tasks benefiting from their ability to capture long-range dependencies with multi-head self-attention modules and multi-layer perceptron. However, calculating global attention brings another disadvantage compared with convolutional neural networks, i.e. requiring much more data and computations to converge, which makes it difficult to generalize well on small datasets, which is common in practical applications. Previous works are either focusing on transferring knowledge from large datasets or adjusting the structure for small datasets. After carefully examining the self-attention modules, we discover that the number of trivial attention weights is far greater than the important ones and the accumulated trivial weights are dominating the attention in Vision Transformers due to their large quantity, which is not handled by the attention itself. This will cover useful non-trivial attention and harm the performance when trivial attention includes more noise, e.g. in shallow layers for some backbones. To solve this issue, we proposed to divide attention weights into trivial and non-trivial ones by thresholds, then Suppressing Accumulated Trivial Attention (SATA) weights by proposed Trivial WeIghts Suppression Transformation (TWIST) to reduce attention noise. Extensive experiments on CIFAR-100 and Tiny-ImageNet datasets show that our suppressing method boosts the accuracy of Vision Transformers by up to 2.3%. Code is available at https://github.com/xiangyu8/SATA.

**********

Cross-Modal Semantic Enhanced Interaction for Image-Sentence Retrieval

Xuri Ge, Fuhai Chen, Songpei Xu, Fuxiang Tao, Joemon M. Jose; Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1022-1031

Image-sentence retrieval has attracted extensive research attention in multimedia and computer vision due to its promising application. The key issue lies in jointly learning the visual and textual representation to accurately estimate their similarity. To this end, the mainstream schema adopts an object-word based attention to calculate their relevance scores and refine their interactive representations with the attention features, which, however, neglects the context of the object representation on the inter-object relationship that matches the predicates in sentences. In this paper, we propose a Cross-modal Semantic Enhanced Interaction method, termed CMSEI for image-sentence retrieval, which correlates the intra- and inter-modal semantics between objects and words. In particular, we first design the intra-modal spatial and semantic graphs based reasoning to enhance the semantic representations of objects guided by the explicit relationships of the objects' spatial positions and their scene graph. Then the visual and textual semantic representations are refined jointly via the inter-modal interactive attention and the cross-modal alignment. To correlate the context of objects with the textual context, we further refine the visual semantic representation via the cross-level object-sentence and word-image based interactive attention. Experimental results on seven standard evaluation metrics show that the proposed CMSEI outperforms the state-of-the-art and the alternative approaches on MS-COCO and Flickr30K benchmarks.

**********

Towards Discriminative and Transferable One-Stage Few-Shot Object Detectors

Karim Guirguis, Mohamed Abdelsamad, George Eskandar, Ahmed Hendawy, Matthias Kayser, Bin Yang, Jürgen Beyerer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3760-3769

Recent object detection models have proved valuable for many robotics and manufacturing tasks, but they require large amounts of annotated data for each new class of objects they are trained for. Few-shot object detection (FSOD) aims to address this problem by learning novel classes given only a few samples of annotated data. While competitive results have been achieved using two-stage FSOD detectors, typically faster one-stage FSODs underperform in comparison. We make the discovery that the large gap in performance between two-stage and one-stage FSODs is mainly due to their weak discriminability, which is explained away by a small post-fusion receptive field and a small number of foreground samples in the loss function. We propose a new one-stage FSOD framework to address these limitations - Few-shot RetinaNet (FSRN). Specifically, we propose: (1) a multi-way support training strategy to augment the number of foreground samples for dense meta-detectors during training, (2) an early multi-level feature fusion providing a wide receptive field that covers the whole anchor area, (3) two augmentation techniques on query and source images to enhance transferability. Extensive experiments demonstrate that the proposed approach addresses the limitations of previous methods and boosts both discriminability and transferability. FSRN is two times faster than twostage FSODs while remaining competitive in accuracy, and it triples the state-of-the-art of one-stage meta-detectors on the competitive 10-shot MS-COCO benchmark. On the PASCAL VOC benchmark, the proposed approach consistently outperforms one-stage meta-detectors and many two-stage FSODs.

**********

LayerDoc: Layer-Wise Extraction of Spatial Hierarchical Structure in Visually-Rich Documents

Puneet Mathur, Rajiv Jain, Ashutosh Mehra, Jiuxiang Gu, Franck Dernoncourt, Anandhavelu N., Quan Tran, Verena Kaynig-Fittkau, Ani Nenkova, Dinesh Manocha, Vlad I. Morariu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3610-3620

Digital documents often contain images and scanned text. Parsing such visually-rich documents is a core task for workflow automation, but it remains challenging since most documents do not encode explicit layout information, e.g., how characters and words are grouped into boxes and ordered into larger semantic entities. Current state-of-the-art layout extraction methods are challenged on such docu

ments as they rely on word sequences to have correct reading order and do not ex
ploit their hierarchical structure. We propose LayerDoc, an approach that uses v
isual features, textual semantics, and spatial coordinates along with constraint
 inference to extract the hierarchical layout structure of documents in a bottom
-up layer-wise fashion. LayerDoc recursively groups smaller regions into larger
semantic elements in 2D to infer complex nested hierarchies. Experiments show th
at our approach outperforms competitive baselines by 10-15% on three diverse dat
asets of forms and mobile app screen layouts for the tasks of spatial region cla
ssification, higher-order group identification, layout hierarchy extraction, rea
ding order detection, and word grouping.
**********

SSSD: Self-Supervised Self Distillation
Wei-Chi Chen, Wei-Ta Chu; Proceedings of the IEEE/CVF Winter Conference on Appli
cations of Computer Vision (WACV), 2023, pp. 2770-2777
With labeled data, self distillation (SD) has been proposed to develop compact b
ut effective models without a complex teacher model available in advance. Such a
pproaches need labeled data to guide the self distillation process. Inspired by
self-supervised (SS) learning, we propose a self-supervised self distillation (S
SSD) approach in this work. Based on an unlabeled image dataset, a model is cons
tructed to learn visual representations in a self-supervised manner. This pre-tr
ained model is then adopted to extract visual representations of the target data
set and generates pseudo labels via clustering. The pseudo labels guide the SD p
rocess, and thus enable SD to proceed in an unsupervised way (no data labels are
 required at all). We verify this idea based on evaluations on the CIFAR-10, CIF
AR-100, and ImageNet-1K datasets, and demonstrate the effectiveness of this unsu
pervised SD approach. Performance outperforming similar frameworks is also shown
.
**********

CellTranspose: Few-Shot Domain Adaptation for Cellular Instance Segmentation
Matthew R. Keaton, Ram J. Zaveri, Gianfranco Doretto; Proceedings of the IEEE/CV
F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 455-466
Automated cellular instance segmentation is a process utilized for accelerating
biological research for the past two decades, and recent advancements have produ
ced higher quality results with less effort from the biologist. Most current end
eavors focus on completely cutting the researcher out of the picture by generati
ng highly generalized models. However, these models invariably fail when faced w
ith novel data, distributed differently than the ones used for training. Rather
than approaching the problem with methods that presume the availability of large
 amounts of target data and computing power for retraining, in this work we addr
ess the even greater challenge of designing an approach that requires minimal am
ounts of new annotated data as well as training time. We do so by designing spec
ialized contrastive losses that leverage the few annotated samples very efficien
tly. A large set of results show that 3 to 5 annotations lead to models with acc
uracy that: 1) significantly mitigate the covariate shift effects; 2) matches or
 surpasses other adaptation methods; 3) even approaches methods that have been f
ully retrained on the target distribution. The adaptation training is only a few
 minutes, paving a path towards a balance between model performance, computing r
equirements and expert-level annotation needs.
**********

Hard To Track Objects With Irregular Motions and Similar Appearances? Make It Ea
sier by Buffering the Matching Space
Fan Yang, Shigeyuki Odashima, Shoichi Masui, Shan Jiang; Proceedings of the IEEE
/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4799
-4808
We propose a Cascaded Buffered IoU (C-BIoU) tracker to track multiple objects th
at have irregular motions and indistinguishable appearances. When appearance fea
tures are unreliable and geometric features are confused by irregular motions, a
pplying conventional Multiple Object Tracking (MOT) methods may generate unsatis
factory results. To address this issue, our C-BIoU tracker adds buffers to expan
d the matching space of detections and tracks, which mitigates the effect of irr

egular motions in two aspects: one is to directly match identical but non-overla
pping detections and tracks in adjacent frames, and the other is to compensate f
or the motion estimation bias in the matching space. In addition, to reduce the
risk of overexpansion of the matching space, cascaded matching is employed: firs
t matching alive tracks and detections with a small buffer, and then matching un
matched tracks and detections with a large buffer. Despite its simplicity, our C
-BIoU tracker works surprisingly well and achieves state-of-the-art results on M
OT datasets that focus on irregular motions and indistinguishable appearances. M
oreover, the C-BIoU tracker is the dominant component for our 2nd place solution
 in the CVPR'22 SoccerNet MOT and the ECCV'22 MOTComplex DanceTrack challenges.
Finally, we analyze the limitation of our C-BIoU tracker in ablation studies and
 discuss its application scope.
**********

Self Supervised Low Dose Computed Tomography Image Denoising Using Invertible Ne
twork Exploiting Inter Slice Congruence

Sutanu Bera, Prabir Kumar Biswas; Proceedings of the IEEE/CVF Winter Conference
on Applications of Computer Vision (WACV), 2023, pp. 5614-5623

The resurgence of deep neural networks has created an alternative pathway for lo
w-dose computed tomography denoising by learning a nonlinear transformation func
tion between low-dose CT (LDCT) and normal-dose CT (NDCT) image pairs. However,
those paired LDCT and NDCT images are rarely available in the clinical environme
nt, making deep neural network deployment infeasible. This study proposes a nove
l method for self-supervised low-dose CT denoising to alleviate the requirement
of paired LDCT and NDCT images. Specifically, we have trained an invertible neur
al network to minimize the pixel-based mean square distance between a noisy slic
e and the average of its two immediate adjacent noisy slices. We have shown the
aforementioned is similar to training a neural network to minimize the distance
between clean NDCT and noisy LDCT image pairs. Again, during the reverse mapping
 of the invertible network, the output image is mapped to the original input ima
ge, similar to cycle consistency loss. Finally, the trained invertible network's
 forward mapping is used for denoising LDCT images. Extensive experiments on two
 publicly available datasets showed that our method performs favourably against
other existing unsupervised methods.
**********

Leveraging Local Patch Differences in Multi-Object Scenes for Generative Adversa
rial Attacks

Abhishek Aich, Shasha Li, Chengyu Song, M. Salman Asif, Srikanth V. Krishnamurth
y, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Appli
cations of Computer Vision (WACV), 2023, pp. 1308-1318

State-of-the-art generative model-based attacks against image classifiers overwh
elmingly focus on single-object (ie., single dominant object) images. Different
from such settings, we tackle a more practical problem of generating adversarial
 perturbations using multi-object (ie., multiple dominant objects) images as the
y are representative of most real-world scenes. Our goal is to design an attack
strategy that can learn from such natural scenes by leveraging the local patch d
ifferences that occur inherently in such images (eg. difference between the loca
l patch on the object 'person' and the object 'bike' in a traffic scene). Our ke
y idea is to misclassify an adversarial multi-object image by confusing the vict
im classifier for each local patch in the image. Based on this, we propose a nov
el generative attack (called Local Patch Difference or LPD-Attack) where a novel
 contrastive loss function uses the aforesaid local differences in feature space
 of multi-object scenes to optimize the perturbation generator. Through various
experiments across diverse victim convolutional neural networks, we show that ou
r approach outperforms baseline generative attacks with highly transferable pert
urbations when evaluated under different white-box and black-box settings.
**********

PRN: Panoptic Refinement Network

Bo Sun, Jason Kuen, Zhe Lin, Philippos Mordohai, Simon Chen; Proceedings of the
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp.
3963-3973

Panoptic segmentation is the task of uniquely assigning every pixel in an image to either a semantic label or an individual object instance, generating a coherent and complete scene description. Many current panoptic segmentation methods, however, predict masks of semantic classes and object instances in separate branches, yielding inconsistent predictions. Moreover, because state-of-the-art panoptic segmentation models rely on box proposals, the instance masks predicted are often of low-resolution. To overcome these limitations, we propose the Panoptic Refinement Network (PRN), which takes masks from base panoptic segmentation models and refines them jointly to produce coherent results. PRN extends the offset map-based architecture of Panoptic-Deeplab with several novel ideas including a foreground mask and instance bounding box offsets, as well as coordinate convolutions for improved spatial prediction. Experimental results on COCO and Cityscapes show that PRN can significantly improve already accurate results from a variety of panoptic segmentation networks.

**********

Controllable 3D Generative Adversarial Face Model via Disentangling Shape and Appearance

Fariborz Taherkhani, Aashish Rai, Quankai Gao, Shaunak Srivastava, Xuanbai Chen, Fernando de la Torre, Steven Song, Aayush Prakash, Daeil Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 826-836

3D face modeling has been an active area of research in computer vision and computer graphics, fueling applications ranging from facial expression transfer in virtual avatars to synthetic data generation. Existing 3D deep learning generative models (e.g., VAE, GANs) allow generating compact face representations (both shape and texture) that can model non-linearities in the shape and appearance space (e.g., scatter effects, specularities,..). However, they lack the capability to control the generation of subtle expressions. This paper proposes a new 3D face generative model that can decouple identity and expression and provides granular control over expressions. In particular, we propose using a pair of supervised auto-encoder and generative adversarial networks to produce high-quality 3D faces, both in terms of appearance and shape. Experimental results in the generation of 3D faces learned with holistic expression labels, or Action Unit (AU) labels, show how we can decouple identity and expression; gaining fine-control over expressions while preserving identity.

**********

Self-Supervised Monocular Depth Estimation: Solving the Edge-Fattening Problem

Xingyu Chen, Ruonan Zhang, Ji Jiang, Yan Wang, Ge Li, Thomas H. Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5776-5786

Self-supervised monocular depth estimation (MDE) models universally suffer from the notorious edge-fattening issue. Triplet loss, popular for metric learning, has made a great success in many computer vision tasks. In this paper, we redesign the patch-based triplet loss in MDE to alleviate the ubiquitous edge-fattening issue. We show two drawbacks of the raw triplet loss in MDE and demonstrate our problem-driven redesigns. First, we present a min. operator based strategy applied to all negative samples, to prevent well-performing negatives sheltering the error of edge-fattening negatives. Second, we split the anchor-positive distance and anchor-negative distance from within the original triplet, which directly optimizes the positives without any mutual effect with the negatives. Extensive experiments show the combination of these two small redesigns can achieve unprecedented results: Our powerful and versatile triplet loss not only makes our model outperform all previous SoTA by a large margin, but also provides substantial performance boosts to a large number of existing models, while introducing no extra inference computation at all.

**********

MonoDVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-Aware Video Panoptic Segmentation

Andra Petrovai, Sergiu Nedevschi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3077-3086

Depth-aware video panoptic segmentation tackles the inverse projection problem o f restoring panoptic 3D point clouds from video sequences, where the 3D points a re augmented with semantic classes and temporally consistent instance identifier s. We propose a novel solution with a multi-task network that performs monocular depth estimation and video panoptic segmentation. Since acquiring ground truth labels for both depth and image segmentation has a relatively large cost, we lev erage the power of unlabeled video sequences with self-supervised monocular dept h estimation and semi-supervised learning from pseudo-labels for video panoptic segmentation. To further improve the depth prediction, we introduce panoptic-gui ded depth losses and a novel panoptic masking scheme for moving objects to avoid corrupting the training signal. Extensive experiments on the Cityscapes-DVPS an d SemKITTI-DVPS datasets demonstrate that our model with the proposed improvemen ts achieves competitive results and fast inference speed.

**********

Wavelength-Aware 2D Convolutions for Hyperspectral Imaging

Leon Amadeus Varga, Martin Messmer, Nuri Benbarka, Andreas Zell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3788-3797

Deep Learning could drastically boost the classification accuracy for Hyperspect ral Imaging (HSI). Still, the training on the mostly small hyperspectral data se ts is not trivial. Two key challenges are the large channel dimension of the rec ordings and the incompatibility between cameras of different manufacturers. By i ntroducing a suitable model bias and continuously defining the channel dimension , we propose a 2D convolution optimized for these challenges of Hyperspectral Im aging. We evaluate the method based on two different hyperspectral applications (inline inspection and remote sensing). Besides the shown superiority of the mod el, the modification adds additional explanatory power. In addition, the model l earns the necessary camera filters in a data-driven manner. Based on these camer a filters, an optimal camera can be designed.

**********

Contrastive Losses Are Natural Criteria for Unsupervised Video Summarization

Zongshang Pang, Yuta Nakashima, Mayu Otani, Hajime Nagahara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2010-2019

Video summarization aims to select a most informative subset of frames in a vide o to facilitate efficient video browsing. Unsupervised methods usually rely on h euristic training objectives such as diversity and representativeness. However, such methods need to bootstrap the online-generated summaries to compute the obj ectives for importance score regression. We consider such a pipeline inefficient and seek to directly quantify the frame-level importance with the help of contr astive losses in the representation learning literature. Leveraging the contrast ive losses, we propose three metrics featuring a desirable key frame: local diss imilarity, global consistency, and uniqueness. With features pre-trained on an i mage classification task, the metrics can already yield high-quality importance scores, demonstrating better or competitive performance compared with past heavi ly-trained methods. We show that by refining the pre-trained features with contr astive learning, the frame-level importance scores can be further improved, and the model can learn from random videos and generalize to test videos with decent performance.

**********

Spatially Multi-Conditional Image Generation

Nikola Popovi■, Ritika Chakraborty, Danda Pani Paudel, Thomas Probst, Luc Van Go ol; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi sion (WACV), 2023, pp. 734-743

In most scenarios, conditional image generation can be thought of as an inversio n of the image understanding process. Since generic image understanding involves solving multiple tasks, it is natural to aim at generating images via multi con ditioning. However, multi-conditional image generation is a very challenging pro blem due to the heterogeneity and the sparsity of the (in practice) available co nditioning labels. In this work, we propose a novel neural architecture to addre

ss the problem of heterogeneity and sparsity of the spatially multi-conditional labels. Our choice of spatial conditioning, such as by semantics and depth, is driven by the promise it holds for better control of the image generation process. The proposed method uses a transformer-like architecture operating pixel-wise, which receives the available labels as input tokens to merge them in a learned homogeneous space of labels. The merged labels are then used for image generation via conditional generative adversarial training. In this process, the sparsity of the labels is handled by simply dropping the input tokens corresponding to the missing labels at the desired locations, thanks to the proposed pixel-wise operating architecture. Our experiments on three benchmark datasets demonstrate the clear superiority of our method over the state-of-the-art and compared baselines.

**********

Towards Online Domain Adaptive Object Detection

Vibashan VS, Poojan Oza, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 478-488

Existing object detection models assume both the training and test data are sampled from the same source domain. This assumption does not hold true when these detectors are deployed in real-world applications, where they encounter new visual domains. Unsupervised Domain Adaptation (UDA) methods are generally employed to mitigate the adverse effects caused by domain shift. Existing UDA methods operate in an offline manner where the model is first adapted toward the target domain and then deployed in real-world applications. However, this offline adaptation strategy is not suitable for real-world applications as the model frequently encounters new domain shifts. Hence, it is critical to develop a feasible UDA method that generalizes to the new domain shifts encountered during deployment time in a continuous online manner. To this end, we propose a novel unified adaptation framework that adapts and improves generalization on the target domain in both offline and online settings. Specifically, we introduce MemXformer - a cross-attention transformer-based memory module where items in the memory take advantage of domain shifts and record prototypical patterns of the target distribution. Further, MemXformer produces strong positive and negative pairs to guide a novel contrastive loss, which enhances target-specific representation learning. Experiments on diverse detection benchmarks show that the proposed strategy producs state-of-the-art performance in both offline and online settings. To the best of our knowledge, this is the first work to address online and offline adaptation settings for object detection. Source code will be released after review.

**********

What Can We Learn by Predicting Accuracy?

Olivier Risser-Maroix, Benjamin Chamand; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2390-2399

This paper seeks to answer the following question: "What can we learn by predicting accuracy?". Indeed, classification is one of the most popular tasks in machine learning, and many loss functions have been developed to maximize this non-differentiable objective function. Unlike past work on loss function design, which was guided mainly by intuition and theory before being validated by experimentation, here we propose to approach this problem in the opposite way: we seek to extract knowledge by experimentation. This data-driven approach is similar to that used in physics to discover general laws from data. We used a symbolic regression method to automatically find a mathematical expression highly correlated with a linear classifier's accuracy. The formula discovered on more than 260 datasets of embeddings has a Pearson's correlation of 0.96 and a r2 of 0.93. More interestingly, this formula is highly explainable and confirms insights from various previous papers on loss design. We hope this work will open new perspectives in the search for new heuristics leading to a deeper understanding of machine learning theory.

**********

nLMVS-Net: Deep Non-Lambertian Multi-View Stereo

Kohei Yamashita, Yuto Enyo, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3037

-3046
We introduce a novel multi-view stereo (MVS) method that can simultaneously reco ver not just per-pixel depth but also surface normals, together with the reflect ance of textureless, complex non-Lambertian surfaces captured under known but na tural illumination. Our key idea is to formulate MVS as an end-to-end learnable network, which we refer to as nLMVS-Net, that seamlessly integrates radiometric cues to leverage surface normals as view-independent surface features for learne d cost volume construction and filtering. It first estimates surface normals as pixel-wise probability densities for each view with a novel shape-from-shading n etwork. These per-pixel surface normal densities and the input multi-view images are then input to a novel cost volume filtering network that learns to recover per-pixel depth and surface normal. The reflectance is also explicitly estimated by alternating with geometry reconstruction. Extensive quantitative evaluations on newly established synthetic and real-world datasets show that nLMVS-Net can robustly and accurately recover the shape and reflectance of complex objects in natural settings.
**********

Ev-NeRF: Event Based Neural Radiance Field
Inwoo Hwang, Junho Kim, Young Min Kim; Proceedings of the IEEE/CVF Winter Confer ence on Applications of Computer Vision (WACV), 2023, pp. 837-847
We present Ev-NeRF, a Neural Radiance Field derived from event data. While event cameras can measure subtle brightness changes in high frame rates, the measurem ents in low lighting or extreme motion suffer from significant domain discrepanc y with complex noise. As a result, the performance of event-based vision tasks d oes not transfer to challenging environments, where the event cameras are expect ed to thrive over normal cameras. We find that the multi-view consistency of NeR F provides a powerful self-supervision signal for eliminating spurious measureme nts and extracting the consistent underlying structure despite highly noisy inpu t. Instead of posed images of the original NeRF, the input to Ev-NeRF is the eve nt measurements accompanied by the movements of the sensors. Using the loss func tion that reflects the measurement model of the sensor, Ev-NeRF creates an integ rated neural volume that summarizes the unstructured and sparse data points capt ured for about 2-4 seconds. The generated neural volume can also produce intensi ty images from novel views with reasonable depth estimates, which can serve as a high-quality input to various vision-based tasks. Our results show that Ev-NeRF achieves competitive performance for intensity image reconstruction under extre me noise conditions and high-dynamic-range imaging.
**********

Jointly Learning Band Selection and Filter Array Design for Hyperspectral Imagin g
Ke Li, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6384-6394
A single-shot multispectral camera equipped with an optimized color filter array (CFA) has the potential to deliver a fast and low-cost hyperspectral (HS) imagi ng system. Previous solutions are largely restricted to designing demosaicing al gorithms for fixed CFAs - be it the Bayer color pattern or evenly-spaced spectra l multiplexing patterns. Since sampling and reconstruction are tightly-coupled, the ability to search for an optimal solution is severely constrained by using p redefined CFAs. In this work, we simultaneously address the problem of spectral band selection, CFA design, image demosaicing, and spectral image recovery in a joint learning framework for single-shot HS imaging. We propose a reinforcement learning (RL) based method for spectral band selection and a novel neural networ k for CFA generation, image demosaicing, and HS image recovery. The final spectr al reconstruction accuracy is used to supervise the training of the main network to maximize the synergies between those tightly-related tasks. The RL method re gards the main network as an agent to collect reward. Our final method delivers a simple setup - as simple as an RGB camera - for HS imaging. Experimental resul ts show that our method outperforms competing methods by a large margin.
**********

InDiReCT: Language-Guided Zero-Shot Deep Metric Learning for Images

Konstantin Kobs, Michael Steininger, Andreas Hotho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1063-1072
Common Deep Metric Learning (DML) datasets specify only one notion of similarity, e.g., two images in the Cars196 dataset are deemed similar if they show the same car model. We argue that depending on the application, users of image retrieval systems have different and changing similarity notions that should be incorporated as easily as possible. Therefore, we present Language-Guided Zero-Shot Deep Metric Learning (LanZ-DML) as a new DML setting in which users control the aspects that should be important for image representations without training data by only using natural language. To this end, we propose InDiReCT (Image representations using Dimensionality Reduction on CLIP embedded Texts), a model for LanZ-DML on images that exclusively uses a few text prompts for training. InDiReCT utilizes CLIP as a fixed feature extractor for images and texts and transfers the variation in text prompt embeddings to the image embedding space. Extensive experiments on five datasets and overall thirteen similarity notions show that, despite not seeing any images during training, InDiReCT performs better than strong baselines and approaches the performance of fully-supervised models. An analysis reveals that InDiReCT learns to focus on regions of the image that correlate with the desired similarity notion, which makes it a fast to train and easy to use method to create custom embedding spaces only using natural language.
**********

Cut-Paste Consistency Learning for Semi-Supervised Lesion Segmentation
Boon Peng Yap, Beng Koon Ng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6160-6169
Semi-supervised learning has the potential to improve the data-efficiency of training data-hungry deep neural networks, which is especially important for medical image analysis tasks where labeled data is scarce. In this work, we present a simple semi-supervised learning method for lesion segmentation tasks based on the ideas of cut-paste augmentation and consistency regularization. By exploiting the mask information available in the labeled data, we synthesize partially labeled samples from the unlabeled images so that the usual supervised learning objective (e.g., binary cross entropy) can be applied. Additionally, we introduce a background consistency term to regularize the training on the unlabeled background regions of the synthetic images. We empirically verify the effectiveness of the proposed method on two public lesion segmentation datasets, including an eye fundus photograph dataset and a brain CT scan dataset. The experiment results indicate that our method achieves consistent and superior performance over other self-training and consistency-based methods without introducing sophisticated network components.
**********

Medical Image Segmentation via Cascaded Attention Decoding
Md Mostafijur Rahman, Radu Marculescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6222-6231
Transformers have shown great promise in medical image segmentation due to their ability to capture long-range dependencies through self-attention. However, they lack the ability to learn the local (contextual) relations among pixels. Previous works try to overcome this problem by embedding convolutional layers either in the encoder or decoder modules of transformers thus ending up sometimes with inconsistent features. To address this issue, we propose a novel attention-based decoder, namely CASCaded Attention DEcoder (CASCADE), which leverages the multi-scale features of hierarchical vision transformers. CASCADE consists of i) an attention gate which fuses features with skip connections and ii) a convolutional attention module that enhances the long-range and local context by suppressing background information. We use a multi-stage feature and loss aggregation framework due to their faster convergence and better performance. Our experiments demonstrate that transformers with CASCADE significantly outperform state-of-the-art CNN- and transformer-based approaches, obtaining up to 5.07% and 6.16% improvements in DICE and mIoU scores, respectively. CASCADE opens new ways of designing better attention-based decoders.
**********

Visualizing Global Explanations of Point Cloud DNNs
Hanxiao Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Co mputer Vision (WACV), 2023, pp. 4741-4750
So far, few researchers have targeted the explainability of point cloud neural n etworks. Part of the explainability methods are not directly applicable to those networks due to the structural specifics. In this work, we show that Activation Maximization (AM) with traditional pixel-wise regularizations fails to generate human-perceptible global explanations for point cloud networks. We propose new generative model-based AM approaches to clearly outline the global explanations and enhance their comprehensibility. Additionally, we propose a composite evalua tion metric to address the limitations of existing evaluating methods, which sim ultaneously takes into account activation value, diversity and perceptibility. E xtensive experiments demonstrate that our generative-based AM approaches outperf orm regularization-based ones both qualitatively and quantitatively. To the best of our knowledge, this is the first work investigating global explainability of point cloud networks. Our code is available at: https://github.com/Explain3D/Po intCloudAM.
**********

LCS: Learning Compressible Subspaces for Efficient, Adaptive, Real-Time Network Compression at Inference Time
Elvis Nunez, Maxwell Horton, Anish Prabhu, Anurag Ranjan, Ali Farhadi, Mohammad Rastegari; Proceedings of the IEEE/CVF Winter Conference on Applications of Comp uter Vision (WACV), 2023, pp. 3818-3827
When deploying deep neural networks (DNNs) to a device, it is traditionally assu med that available computational resources (compute, memory, and power) remain s tatic. However, real-world computing systems do not always provide stable resour ce guarantees. Computational resources need to be conserved when load from other processes is high, or available memory is low. In this work, we present a train ing procedure to produce DNNs that can be compressed in real-time to arbitrary c ompression levels entirely on-device. This enables the deployment of a single mo del that can efficiently adapt to its host device's available resources. We form ulate this problem as learning an adaptively compressible network subspace, wher e one end is optimized for accuracy, and the other for efficiency. Our subspace model requires no recalibration nor retraining when changing compression levels. Moreover, our generic training framework is amenable to multiple forms of compr ession, and we present results for unstructured sparsity, structured sparsity, a nd quantization on a variety of architectures. We present models that require a single extra copy of network parameters, as well as models that require no extra parameters. Both models allow for operation at any compression level within a w ide range (for example, 0% to 90% for structured sparsity with ResNet18 on Image Net). At each compression level, our models achieve an accuracy comparable to a baseline model optimized for that particular compression level. To our knowledge , our method is the first to enable adaptive on-device network compression with little to no computational overhead.
**********

Fine-Context Shadow Detection Using Shadow Removal
Jeya Maria Jose Valanarasu, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1705-1714
Current shadow detection methods perform poorly when detecting shadow regions th at are small, unclear or have blurry edges. In this work, we attempt to address this problem on two fronts. First, we propose a Fine Context-aware Shadow Detect ion Network (FCSD-Net), where we constraint the receptive field size and focus o n low-level features to learn fine context features better. Second, we propose a new learning strategy, called Restore to Detect (R2D), where we show that when a deep neural network is trained for restoration (shadow removal), it learns mea ningful features to delineate the shadow masks as well. To make use of this comp lementary nature of shadow detection and removal tasks, we train an auxiliary ne twork for shadow removal and propose a complementary feature learning block (CFL ) to learn and fuse meaningful features from shadow removal network to the shado w detection network. We train the proposed network, FCSD-Net, using the R2D lear

ning strategy across multiple datasets. Experimental results on three public sha dow detection datasets (ISTD, SBU and UCF) show that our method improves the sha dow detection performance while being able to detect fine context better compare d to the other recent methods. Our proposed learning strategy can also be adopte d easily as a useful pipeline in future advances in shadow detection and removal .

**********

Spatial Consistency Loss for Training Multi-Label Classifiers From Single-Label Annotations

Thomas Verelst, Paul K. Rubenstein, Marcin Eichner, Tinne Tuytelaars, Maxim Berm an; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi sion (WACV), 2023, pp. 3879-3889

Multi-label image classification is more applicable 'in the wild' than single-la bel classification, as natural images usually contain multiple objects. However, exhaustively annotating images with every object of interest is costly and time -consuming. We train multi-label classifiers from datasets where each image is a nnotated with a single positive label only. As the presence of all other classes is unknown, we propose an Expected Negative loss that builds a set of expected negative labels in addition to the annotated positives. This set is determined b ased on prediction consistency, by averaging predictions over consecutive traini ng epochs to build robust targets. Moreover, the crop data-augmentation leads to additional label noise by cropping out the single annotated object. Our novel s patial consistency loss improves supervision and ensures consistency of the spat ial feature maps by maintaining per-class running-average heatmaps for each trai ning image. We use MS-COCO, Pascal VOC, NUS-WIDE and CUB-Birds datasets to demon strate the gains of the Expected Negative loss in combination with consistency a nd spatial consistency losses. We also demonstrate improved multi-label classifi cation mAP on ImageNet-1K using the ReaL multi-label validation set.

**********

ARUBA: An Architecture-Agnostic Balanced Loss for Aerial Object Detection

Rebbapragada V. C. Sairam, Monish Keswani, Uttaran Sinha, Nishit Shah, Vineeth N . Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3719-3728

Deep neural networks tend to reciprocate the bias of their training dataset. In object detection, the bias exists in the form of various imbalances such as clas s, background-foreground, and object size. In this paper, we denote size of an o bject as the number of pixels it covers in an image and size imbalance as the ov er-representation of certain sizes of objects in a dataset. We aim to address th e problem of size imbalance in drone-based aerial image datasets. Existing metho ds for solving size imbalance are based on architectural changes that utilize mu ltiple scales of images or feature maps for detecting objects of different sizes . We, on the other hand, propose a novel ARchitectUre-agnostic BAlanced Loss (AR UBA) that can be applied as a plugin on top of any object detection model. It fo llows a neighborhood-driven approach inspired by the ordinality of object size. We evaluate the effectiveness of our approach through comprehensive experiments on aerial datasets such as HRSC2016, DOTAv1.0, DOTAv1.5 and VisDrone and obtain consistent improvement in performance.

**********

Multimodal Multi-Head Convolutional Attention With Various Kernel Sizes for Medi cal Image Super-Resolution

Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Andreea-Iuliana Miron, Olivian Sa vencu, Nicolae-C■t■lin Ristea, Nicolae Verga, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2195-2205

Super-resolving medical images can help physicians in providing more accurate di agnostics. In many situations, computed tomography (CT) or magnetic resonance im aging (MRI) techniques capture several scans (modes) during a single investigati on, which can jointly be used (in a multimodal fashion) to further boost the qua lity of super-resolution results. To this end, we propose a novel multimodal mul ti-head convolutional attention module to super-resolve CT and MRI scans. Our at

tention module uses the convolution operation to perform joint spatial-channel a
ttention on multiple concatenated input tensors, where the kernel (receptive fie
ld) size controls the reduction rate of the spatial attention, and the number of
 convolutional filters controls the reduction rate of the channel attention, res
pectively. We introduce multiple attention heads, each head having a distinct re
ceptive field size corresponding to a particular reduction rate for the spatial
attention. We integrate our multimodal multi-head convolutional attention (MMHCA
) into two deep neural architectures for super-resolution and conduct experiment
s on three data sets. Our empirical results show the superiority of our attentio
n module over the state-of-the-art attention mechanisms used in super-resolution
. Moreover, we conduct an ablation study to assess the impact of the components
involved in our attention module, e.g. the number of inputs or the number of hea
ds. Our code is freely available at https://github.com/lilygeorgescu/MHCA.
**********

FUSSL: Fuzzy Uncertain Self Supervised Learning
Salman Mohamadi, Gianfranco Doretto, Donald A. Adjeroh; Proceedings of the IEEE/
CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2799-
2808
Self supervised learning (SSL) has become a very successful technique to harness
 the power of unlabeled data, with no annotation effort. A number of developed a
pproaches are evolving with the goal of outperforming supervised alternatives, w
hich have been relatively successful. Similar to some other disciplines in deep
representation learning, one main issue in SSL is robustness of the approaches u
nder different settings. In this paper, for the first time, we recognize the fun
damental limits of SSL coming from the use of a single-supervisory signal. To ad
dress this limitation, we leverage the power of uncertainty representation to de
vise a robust and general standard hierarchical learning/training protocol for a
ny SSL baseline, regard- less of their assumptions and approaches. Essentially,
using the information bottleneck principle, we decompose feature learning into a
 two-stage training procedure, each with a distinct supervision signal. This dou
ble supervision approach is captured in two key steps: 1) invariance enforcement
 to data augmentation, and 2) fuzzy pseudo labeling (both hard and soft annotati
on). This simple, yet, effective protocol which enables cross-class/cluster feat
ure learning, is instantiated via an initial training of an ensemble of models t
hrough invariance enforcement to data augmentation as first training phase, and
then assigning fuzzy labels to the original samples for the second training phas
e. We consider multiple alternative scenarios with double supervision and evalua
te the effectiveness of our approach on recent baselines, covering four differen
t SSL paradigms, including geometrical, contrastive, non-contrastive, and hard/s
oft whitening (redundancy reduction) baselines. We performed extensive experimen
ts under multiple settings to show that the proposed training protocol consisten
tly improves the performance of the former baselines, independent of their respe
ctive underlying principles.
**********

DDNeRF: Depth Distribution Neural Radiance Fields
David Dadon, Ohad Fried, Yacov Hel-Or; Proceedings of the IEEE/CVF Winter Confer
ence on Applications of Computer Vision (WACV), 2023, pp. 755-763
The field of implicit neural representation has made significant progress. Model
s such as neural radiance fields (NeRF), which uses relatively small neural netw
orks, can represent high-quality scenes and achieve state-of-the-art results for
 novel view synthesis. Training these types of networks, however, is still compu
tationally expensive and the model struggles with real life 360 degree scenes. I
n this work, we propose the depth distribution neural radiance field (DDNeRF), a
 new method that significantly increases sampling efficiency along rays during t
raining, while achieving superior results for a given sampling budget. DDNeRF ac
hieves this performance by learning a more accurate representation of the densit
y distribution along rays. More specifically, the proposed framework trains a co
arse model to predict the internal distribution of the transparency of an input
volume along each ray. This estimated distribution then guides the sampling proc
edure of the fine model. Our method allows using fewer samples during training w

hile achieving better output quality with the same computational resources.
**********

Deep Model-Based Super-Resolution With Non-Uniform Blur
Charles Laroche, Andrés Almansa, Matias Tassano; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1797-1808
We propose a state-of-the-art method for super-resolution with non-uniform blur. Single-image super-resolution methods seek to restore a high-resolution image from blurred, subsampled, and noisy measurements. Despite their impressive performance, existing techniques usually assume a uniform blur kernel. Hence, these techniques do not generalize well to the more general case of non-uniform blur. Instead, in this paper, we address the more realistic and computationally challenging case of spatially-varying blur. To this end, we first propose a fast deep plug-and-play algorithm, based on linearized ADMM splitting techniques, which can solve the super-resolution problem with spatially-varying blur. Second, we unfold our iterative algorithm into a single network and train it end-to-end. In this way, we overcome the intricacy of manually tuning the parameters involved in the optimization scheme. Our algorithm presents remarkable performance and generalizes well after a single training to a large family of spatially-varying blur kernels, noise levels and scale factors.
**********

Progressive Video Summarization via Multimodal Self-Supervised Learning
Haopeng Li, Qiuhong Ke, Mingming Gong, Tom Drummond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5584-5593
Modern video summarization methods are based on deep neural networks that require a large amount of annotated data for training. However, existing datasets for video summarization are small-scale, easily leading to over-fitting of the deep models. Considering that the annotation of large-scale datasets is time-consuming, we propose a multimodal self-supervised learning framework to obtain semantic representations of videos, which benefits the video summarization task. Specifically, the self-supervised learning is conducted by exploring the semantic consistency between the videos and text in both course-grained and fine-grained fashions, as well as recovering masked frames in the videos. The multimodal framework is trained on a newly-collected dataset that consists of video-text pairs. Additionally, we introduce a progressive video summarization method, where the important content in a video is pinpointed progressively to generate better summaries. Extensive experiments have proved the effectiveness and superiority of our method in rank correlation coefficients and F-score compared to the state of the art.
**********

Pushing the Efficiency Limit Using Structured Sparse Convolutions
Vinay Kumar Verma, Nikhil Mehta, Shijing Si, Ricardo Henao, Lawrence Carin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6503-6513
Weight pruning is among the most popular approaches for compressing deep convolutional neural networks. Recent work suggests that in a randomly initialized deep neural network, there exist sparse subnetworks that achieve performance comparable to the original network. Unfortunately, finding these subnetworks involves iterative stages of training and pruning, which can be computationally expensive. We propose Structured Sparse Convolution (SSC), that leverages the inherent structure in images to reduce the parameters in the convolutional filter. This leads to improved efficiency of convolutional architectures compared to existing methods that perform pruning at initialization. We show that SSC is a generalization of commonly used layers (depthwise, groupwise, and pointwise convolution) in "efficient architectures." Extensive experiments on well-known CNN models and datasets show the effectiveness of the proposed method. Architectures based on SSC achieve state-of-the-art performance compared to baselines on CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet classification benchmarks.
**********

Robust Real-World Image Enhancement Based on Multi-Exposure LDR Images

Haoyu Ren, Yi Fan, Stephen Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1715-1723
Robust real-world image enhancement from multi-exposure low dynamic range (LDR) images is a challenging task due to the unexpected inconsistency among the input images, such as the large motion or various exposures. In this paper, we propose a novel end-to-end image enhancement network to solve this problem. After extracting contextual information from the LDR images, we design a novel matching volume to align them by considering the motion and exposure differences among the input images. A stacked hourglass with dilated convolution is further utilized to aggregate the matched feature maps to the final enhanced image. In addition, we design a weakly-supervised pairwise loss function to evaluate the color consistency in the enhanced image, which further boosts the performance. We show the effectiveness of our methods on high dynamic ranging imaging (HDR) and End-to-End image signal processing (E2E-ISP). Experimental results demonstrate that our model achieves state-of-the-art enhancement performance.
**********

HOOT: Heavy Occlusions in Object Tracking Benchmark
Gozde Sahin, Laurent Itti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4830-4839
In this paper, we present HOOT, the Heavy Occlusions in Object Tracking Benchmark, a new visual object tracking dataset aimed towards handling high occlusion scenarios for single-object tracking tasks. The benchmark consists of 581 high-quality videos, which have 436K frames densely annotated with rotated bounding boxes for the targets spanning 74 object classes. The dataset is geared for development, evaluation and analysis of visual tracking algorithms that are robust to occlusions. It is comprised of videos with high occlusion levels, where the median percentage of occluded frames per-video is 68%. It also provides critical attributes on occlusions, which include defining a taxonomy for occluders, providing occlusion masks for every bounding box, per-frame partial/full occlusion labels and more. HOOT has been compiled to encourage development of new methods targeting occlusion handling in visual tracking, by providing training and test splits with high occlusion levels. This makes HOOT the first densely-annotated, large dataset designed for single-object tracking under severe occlusion. We evaluate 15 state-of-the-art trackers on this new dataset to act as a baseline for future work focusing on occlusions.
**********

Self-Attentive Pooling for Efficient Deep Learning
Fang Chen, Gourav Datta, Souvik Kundu, Peter A. Beerel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3974-3983
Efficient custom pooling techniques that can aggressively trim the dimensions of a feature map for resource-constrained computer vision applications have recently gained significant traction. However, prior pooling works extract only the local context of the activation maps, limiting their effectiveness. In contrast, we propose a novel non-local self-attentive pooling method that can be used as a drop-in replacement to the standard pooling layers, such as max/average pooling or strided convolution. The proposed self-attention module uses patch embedding, multi-head self-attention, and spatial-channel restoration, followed by sigmoid activation and exponential soft-max. This self-attention mechanism efficiently aggregates dependencies between non-local activation patches during down-sampling. Extensive experiments on standard object classification and detection tasks with various convolutional neural network (CNN) architectures demonstrate the superiority of our proposed mechanism over the state-of-the-art (SOTA) pooling techniques. In particular, we surpass the test accuracy of existing pooling techniques on different variants of MobileNet-V2 on ImageNet by an average of 1.2%. With the aggressive down-sampling of the activation maps in the initial layers (providing up to 22x reduction in memory consumption), our approach achieves 1.43% higher test accuracy compared to SOTA techniques with iso-memory footprints. This enables the deployment of our models in memory-constrained devices, such as micro-controllers without losing significant accuracy, because the initial activatio

n maps consume a significant amount of on-chip memory for high-resolution images required for complex vision tasks. Our pooling method also leverages channel pruning to further reduce memory footprints. Codes are available at https://github.com/C-Fun/Non-Local-Pooling.

**********

Self-Distilled Self-Supervised Representation Learning

Jiho Jang, Seonhoon Kim, Kiyoon Yoo, Chaerin Kong, Jangho Kim, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2829-2839

State-of-the-art frameworks in self-supervised learning have recently shown that fully utilizing transformer-based models can lead to performance boost compared to conventional CNN models. Striving to maximize the mutual information of two views of an image, existing works apply a contrastive loss to the final representations. Motivated by self-distillation in the supervised regime, we further exploit this by allowing the intermediate representations to learn from the final layer via the contrastive loss. Through self-distillation, the intermediate layers are better suited for instance discrimination, making the performance of an early-exited sub-network not much degraded from that of the full network. This renders the pretext task easier also for the final layer, lead to better representations. Our method, Self-Distilled Self-Supervised Learning (SDSSL), outperforms competitive baselines (SimCLR, BYOL and MoCo v3) using ViT on various tasks and datasets. In the linear evaluation and k-NN protocol, SDSSL not only leads to superior performance in the final layers, but also in most of the lower layers. Furthermore, qualitative and quantative analyses show how representations are formed more effectively along the transformer layers. Code will be available.

**********

Composite Learning for Robust and Effective Dense Predictions

Menelaos Kanakis, Thomas E. Huang, David Brüggemann, Fisher Yu, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2299-2308

Multi-task learning promises better model generalization on a target task by jointly optimizing it with an auxiliary task. However, the current practice requires additional labeling efforts for the auxiliary task, while not guaranteeing better model performance. In this paper, we find that jointly training a dense prediction (target) task with a self-supervised (auxiliary) task can consistently improve the performance of the target task, while eliminating the need for labeling auxiliary tasks. We refer to this joint training as Composite Learning (CompL). Experiments of CompL on monocular depth estimation, semantic segmentation, and boundary detection show consistent performance improvements in fully and partially labeled datasets. Further analysis on depth estimation reveals that joint training with self-supervision outperforms most labeled auxiliary tasks. We also find that CompL can improve model robustness when the models are evaluated in new domains. These results demonstrate the benefits of self-supervision as an auxiliary task, and establish the design of novel task-specific self-supervised methods as a new axis of investigation for future multi-task learning research.

**********

Efficient Skeleton-Based Action Recognition via Joint-Mapping Strategies

Min-Seok Kang, Dongoh Kang, HanSaem Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3403-3412

Graph convolutional networks (GCNs) have brought remarkable progress in skeleton-based action recognition. However, high computational cost and large model size make models difficult to be applied in real-world embedded system. Specifically, GCN that is applied in automated surveillance system pre-require models such as pedestrian detection and human pose estimation. Therefore, each model should be computationally lightweight and whole process should be operated in real-time. In this paper, we propose two different joint-mapping modules to reduce the number of joint representations, alleviating a total computational cost and model size. Our models achieve better accuracy-latency trade-off compared to previous state-of-the-arts on two datasets, namely NTU RGB+D and NTU RGB+D 120, demonstrating the suitability for practical applications. Furthermore, we measure the late

ncy of the models by using TensorRT framework to compare the models from a practical perspective.
**********

PointInverter: Point Cloud Reconstruction and Editing via a Generative Model With Shape Priors
Jaeyeon Kim, Binh-Son Hua, Thanh Nguyen, Sai-Kit Yeung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 592-601
In this paper, we propose a new method for mapping a 3D point cloud to the latent space of a 3D generative adversarial network. Our generative model for 3D point clouds is based on SP-GAN, a state-of-the-art sphere-guided 3D point cloud generator. We derive an efficient way to encode an input 3D point cloud to the latent space of the SP-GAN. Our point cloud encoder can resolve the point ordering issue during inversion, and thus can determine the correspondences between points in the generated 3D point cloud and those in the canonical sphere used by the generator. We show that our method outperforms previous GAN inversion methods for 3D point clouds, achieving the state-of-the-art results both quantitatively and qualitatively. Our code is available upon publication.
**********

AdvisIL - A Class-Incremental Learning Advisor
Eva Feillet, Grégoire Petit, Adrian Popescu, Marina Reyboz, Céline Hudelot; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2400-2409
Recent class-incremental learning methods combine deep neural architectures and learning algorithms to handle streaming data under memory and computational constraints. The performance of existing methods varies depending on the characteristics of the incremental process. To date, there is no other approach than to test all pairs of learning algorithms and neural architectures on the training data available at the start of the learning process to select a suited algorithm-architecture combination. To tackle this problem, in this article, we introduce AdvisIL, a method which takes as input the main characteristics of the incremental process (memory budget for the deep model, initial number of classes, size of incremental steps) and recommends an adapted pair of learning algorithm and neural architecture. The recommendation is based on a similarity between the user-provided settings and a large set of pre-computed experiments. AdvisIL makes class-incremental learning easier, since users do not need to run cumbersome experiments to design their system. We evaluate our method on four datasets under six incremental settings and three deep model sizes. We compare six algorithms and three deep neural architectures. Results show that AdvisIL has better overall performance than any of the individual combinations of a learning algorithm and a neural architecture. AdvisIL's code is available at https://github.com/EvaJF/AdvisIL.
**********

LAB: Learnable Activation Binarizer for Binary Neural Networks
Sieger Falkena, Hadi Jamali-Rad, Jan van Gemert; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6425-6434
Binary Neural Networks (BNNs) are receiving an upsurge of attention for bringing power-hungry deep learning towards edge devices. The traditional wisdom in this space is to employ sign() for binarizing featuremaps. We argue and illustrate that sign() is a uniqueness bottleneck, limiting information propagation throughout the network. To alleviate this, we propose to dispense sign(), replacing it with a learnable activation binarizer (LAB), allowing the network to learn a fine-grained binarization kernel per layer - as opposed to global thresholding. LAB is a novel universal module that can seamlessly be integrated into existing architectures. To confirm this, we plug it into four seminal BNNs and show a considerable performance boost at the cost of tolerable increase in delay and complexity. Finally, we build an end-to-end BNN (coined as LAB-BNN) around LAB, and demonstrate that it achieves competitive performance on par with the state-of-the-art on ImageNet. Codebase in the supplementary will be made publicly available upon acceptance.
**********

Fine-Grained Activities of People Worldwide
Jeffrey Byrne, Gregory Castañón, Zhongheng Li, Gil Ettinger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3308-3319
Every day, humans perform many closely related activities that involve subtle discriminative motions, such as putting on a shirt vs. putting on a jacket, or shaking hands vs. giving a high five. Activity recognition by ethical visual AI could provide insights into our patterns of daily life, however existing activity recognition datasets do not capture the massive diversity of these human activities around the world. To address this limitation, we introduce Collector, a free mobile app to record video while simultaneously annotating objects and activities of consented subjects. This new data collection platform was used to curate the Consented Activities of People (CAP) dataset, the first large-scale, fine-grained activity dataset of people worldwide. The CAP dataset contains 1.45M video clips of 512 fine grained activity labels of daily life, grouped into 144 coarse activity classes, collected by 780 subjects in 33 countries. We provide activity classification and activity detection benchmarks for this dataset, and analyze baseline results to gain insight into how people around with world perform common activities. The dataset, benchmarks, evaluation tools, public leaderboards and mobile apps are available for use at visym.github.io/cap.
**********
Recur, Attend or Convolve? On Whether Temporal Modeling Matters for Cross-Domain Robustness in Action Recognition
Sofia Broomé, Ernest Pokropek, Boyu Li, Hedvig Kjellström; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4199-4209
Most action recognition models today are highly parameterized, and evaluated on datasets with appearance-wise distinct classes. It has also been shown that 2D Convolutional Neural Networks (CNNs) tend to be biased toward texture rather than shape in still image recognition tasks, in contrast to humans. Taken together, this raises suspicion that large video models partly learn spurious spatial texture correlations rather than to track relevant shapes over time to infer generalizable semantics from their movement. A natural way to avoid parameter explosion when learning visual patterns over time is to make use of recurrence. Biological vision consists of abundant recurrent circuitry, and is superior to computer vision in terms of domain shift generalization. In this article, we empirically study whether the choice of low-level temporal modeling has consequences for texture bias and cross-domain robustness. In order to enable a light-weight and systematic assessment of the ability to capture temporal structure, not revealed from single frames, we provide the Temporal Shape (TS) dataset, as well as modified domains of Diving48 allowing for the investigation of spatial texture bias in video models. The combined results of our experiments indicate that sound physical inductive bias such as recurrence in temporal modeling may be advantageous when robustness to domain shift is important for the task.
**********
Rethinking Rotation in Self-Supervised Contrastive Learning: Adaptive Positive or Negative Data Augmentation
Atsuyuki Miyai, Qing Yu, Daiki Ikami, Go Irie, Kiyoharu Aizawa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2809-2818
Rotation is frequently listed as a candidate for data augmentation in contrastive learning but seldom provides satisfactory improvements. We argue that this is because the rotated image is always treated as either positive or negative. The semantics of an image can be rotation-invariant or rotation-variant, so whether the rotated image is treated as positive or negative should be determined based on the content of the image. Therefore, we propose a novel augmentation strategy, adaptive Positive or Negative Data Augmentation (PNDA), in which an original and its rotated image are a positive pair if they are semantically close and a negative pair if they are semantically different. To achieve PNDA, we first determine whether rotation is positive or negative on an image-by-image basis in an un

supervised way. Then, we apply PNDA to contrastive learning frameworks. Our experiments showed that PNDA improves the performance of contrastive learning.
**********

Far3Det: Towards Far-Field 3D Detection
Shubham Gupta, Jeet Kanjani, Mengtian Li, Francesco Ferroni, James Hays, Deva Ramanan, Shu Kong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 692-701
We focus on the task of far-field 3D detection (Far3Det) of objects beyond a certain distance from an observer, e.g., >50m. Far3Det is particularly important for autonomous vehicles (AVs) operating at highway speeds, which require detections of far-field obstacles to ensure sufficient braking distances. However, contemporary AV benchmarks such as nuScenes underemphasize this problem because they evaluate performance only up to a certain distance (50m). One reason is that obtaining far-field 3D annotations is difficult, particularly for lidar sensors that produce very few point returns for far-away objects. Indeed, we find that almost 50% of far-field objects (beyond 50m) contain zero lidar points. Secondly, current metrics for 3D detection employ a "one-size-fits-all" philosophy, using the same tolerance thresholds for near and far objects, inconsistent with tolerances for both human vision and stereo disparities. Both factors lead to an incomplete analysis of the Far3Det task. For example, while conventional wisdom tells us that high-resolution RGB sensors should be vital for the 3D detection of far-away objects, lidar-based methods still rank higher compared to RGB counterparts on the current benchmark leaderboards. As a first step towards a Far3Det benchmark, we develop a method to find well-annotated scenes from the nuScenes dataset and derive a well-annotated far-field validation set. We also propose a Far3Det evaluation protocol and explore various 3D detection methods for Far3Det. Our result convincingly justifies the long held conventional wisdom that high-resolution RGB improves 3D detection in the far-field. We further propose a simple yet effective method that fuses detections from RGB and lidar detectors based on non-maximum suppression, which remarkably outperforms state-of-the-art 3D detectors in the far-field.
**********

Towards MOOCs for Lipreading: Using Synthetic Talking Heads To Train Humans in Lipreading at Scale
Aditya Agarwal, Bipasha Sen, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2217-2226
Many people with some form of hearing loss consider lipreading as their primary mode of day-to-day communication. However, finding resources to learn or improve one's lipreading skills can be challenging. This is further exacerbated in the COVID19 pandemic due to restrictions on direct interactions with peers and speech therapists. Today, online MOOCs platforms like Coursera and Udemy have become the most effective form of training for many types of skill development. However, online lipreading resources are scarce as creating such resources is an extensive process needing months of manual effort to record hired actors. Because of the manual pipeline, such platforms are also limited in vocabulary, supported languages, accents, and speakers and have a high usage cost. In this work, we investigate the possibility of replacing real human talking videos with synthetically generated videos. Synthetic data can easily incorporate larger vocabularies, variations in accent, and even local languages and many speakers. We propose an end-to-end automated pipeline to develop such a platform using state-of-the-art talking head video generator networks, text-to-speech models, and computer vision techniques. We then perform an extensive human evaluation using carefully thought out lipreading exercises to validate the quality of our designed platform against the existing lipreading platforms. Our studies concretely point toward the potential of our approach in developing a large-scale lipreading MOOC platform that can impact millions of people with hearing loss.
**********

Modality Mixer for Multi-Modal Action Recognition
Sumin Lee, Sangmin Woo, Yeonju Park, Muhammad Adi Nugroho, Changick Kim; Proceed

ings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3298-3307
In multi-modal action recognition, it is important to consider not only the complementary nature of different modalities but also global action content. In this paper, we propose a novel network, named Modality Mixer (M-Mixer) network, to leverage complementary information across modalities and temporal context of an action for multi-modal action recognition. We also introduce a simple yet effective recurrent unit, called Multi-modal Contextualization Unit (MCU), which is a core component of M-Mixer. Our MCU temporally encodes a sequence of one modality (e.g., RGB) with action content features of other modalities (e.g., depth, IR). This process encourages M-Mixer to exploit global action content and also to supplement complementary information of other modalities. As a result, our proposed method outperforms state-of-the-art methods on NTU RGB+D 60, NTU RGB+D 120, and NWUCLA datasets. Moreover, we demonstrate the effectiveness of M-Mixer by conducting comprehensive ablation studies.
**********
Relaxing Contrastiveness in Multimodal Representation Learning
Zudi Lin, Erhan Bas, Kunwar Yashraj Singh, Gurumurthy Swaminathan, Rahul Bhotika; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2227-2236
Multimodal representation learning for images with paired raw texts can improve the usability and generality of the learned semantic concepts while significantly reducing annotation costs. In this paper, we explore the design space of loss functions in visual-linguistic pretraining frameworks and propose a novel Relaxed Contrastive (ReCo) objective, which acts as a drop-in replacement of the widely used InfoNCE loss. The key insight of ReCo is to allow a relaxed negative space by not penalizing unpaired multimodal samples (ie, negative pairs) that are already orthogonal or negatively correlated. Unlike the widely-used InfoNCE, which keeps repelling negative pairs as long as they are not anti-correlated, ReCo by design embraces more diversity and flexibility of the learned embeddings. We conduct extensive experiments using ReCo with state-of-the-art models by pretraining on the MIMIC-CXR dataset that consists of chest radiographs and free-text radiology reports, and evaluating on the CheXpert dataset for multimodal retrieval and disease classification. Our ReCo achieves an absolute improvement of 2.9% over the InfoNCE baseline on the CheXpert Retrieval dataset in average retrieval precision and reports better or comparable performance in the linear evaluation and finetuning for classification. We further show that ReCo outperforms InfoNCE on the Flickr30K dataset by 1.7% in retrieval Recall@1, demonstrating the generalizability of our approach to natural images.
**********
Towards Disturbance-Free Visual Mobile Manipulation
Tianwei Ni, Kiana Ehsani, Luca Weihs, Jordi Salvador; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5219-5231
Deep reinforcement learning has shown promising results on an abundance of robotic tasks in simulation, including visual navigation and manipulation. Prior work generally aims to build embodied agents that solve their assigned tasks as quickly as possible, while largely ignoring the problems caused by collision with objects during interaction. This lack of prioritization is understandable: there is no inherent cost in breaking virtual objects. As a result, "well-trained" agents frequently collide with objects before achieving their primary goals, a behavior that would be catastrophic in the real world. In this paper, we study the problem of training agents to complete the task of visual mobile manipulation in the ManipulaTHOR environment while avoiding unnecessary collision (disturbance) with objects. We formulate disturbance avoidance as a penalty term in the reward function, but find that directly training with such penalized rewards often results in agents being unable to escape poor local optima. Instead, we propose a two-stage training curriculum where an agent is first allowed to freely explore and build basic competencies without penalization, after which a disturbance penalty is introduced to refine the agent's behavior. Results on testing scenes show

that our curriculum not only avoids these poor local optima, but also leads to 10% absolute gains in success rate without disturbance, compared to our state-of-the-art baselines. Moreover, our curriculum is significantly more performant than a safe RL algorithm that casts collision avoidance as a constraint. Finally, we propose a novel disturbance-prediction auxiliary task that accelerates learning.

**********

Exploiting Instance-Based Mixed Sampling via Auxiliary Source Domain Supervision for Domain-Adaptive Action Detection

Yifan Lu, Gurkirt Singh, Suman Saha, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4145-4156

We propose a novel domain adaptive action detection approach and a new adaptation protocol that leverages the recent advancements in image-level unsupervised domain adaptation (UDA) techniques and handle vagaries of instance-level video data. Self-training combined with cross-domain mixed sampling has shown remarkable performance gain in semantic segmentation in UDA (unsupervised domain adaptation) context. Motivated by this fact, we propose an approach for human action detection in videos that transfers knowledge from the source domain (annotated dataset) to the target domain (unannotated dataset) using mixed sampling and pseudo-label-based selftraining. The existing UDA techniques follow a ClassMix algorithm for semantic segmentation. However, simply adopting ClassMix for action detection does not work, mainly because these are two entirely different problems, i.e., pixel-label classification vs. instance-label detection. To tackle this, we propose a novel action instance mixed sampling technique that combines information across domains based on action instances instead of action classes. Moreover, we propose a new UDA training protocol that addresses the long-tail sample distribution and domain shift problem by using supervision from an auxiliary source domain (ASD). For the ASD, we propose a new action detection dataset with dense frame-level annotations. We name our proposed framework as domain-adaptive action instance mixing (DA-AIM). We demonstrate that DA-AIM consistently outperforms prior works on challenging domain adaptation benchmarks. The source code is available at https://github.com/wwwfan628/DA-AIM.

**********

Graph-Based Self-Learning for Robust Person Re-Identification

Yuqiao Xian, Jinrui Yang, Fufu Yu, Jun Zhang, Xing Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4789-4798

Existing deep learning approaches for person re-identification (Re-ID) mostly rely on large-scale and well-annotated training data. However, human-annotated labels are prone to label noise in real-world applications. Previous person Re-ID works mainly focus on random label noise, which doesn't properly reflect the characteristic of label noise in practical human-annotated process. In this work, we find the visual ambiguity noise is more common and reasonable noise assumption in annotation of person Re-ID. To handle the kind of noise, we propose a simple and effective robust person Re-ID framework, namely Graph-Based Self-Learning (GBSL), to iteratively learn discriminative representation and rectify noisy labels with limited annotated samples for each identity. Meanwhile, considering the practical annotation process in person Re-ID, we further extend the visual ambiguity noise assumption and propose a type of more practical label noise in person Re-ID, namely the tracklet-level label noise (TLN). Without modifying network architecture or loss function, our approach significantly improves the robustness against label noise of the Re-ID system. Our model obtains competitive performance with training data corrupted by various types of label noise and outperforms the existing methods for robust Re-ID on public benchmarks.

**********

SVD-NAS: Coupling Low-Rank Approximation and Neural Architecture Search

Zhewen Yu, Christos-Savvas Bouganis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1503-1512

The task of compressing pre-trained Deep Neural Networks has attracted wide interest of the research community due to its great benefits in freeing practitioner

s from data access requirements. In this domain, low-rank approximation is a pro mising method, but existing solutions considered a restricted number of design c hoices and failed to efficiently explore the design space, which lead to severe accuracy degradation and limited compression ratio achieved. To address the abov e limitations, this work proposes the SVD-NAS framework that couples the domains of low-rank approximation and neural architecture search. SVD-NAS generalises a nd expands the design choices of previous works by introducing the Low-Rank arch itecture space, LR-space, which is a more fine-grained design space of low-rank approximation. Afterwards, this work proposes a gradient-descent-based search fo r efficiently traversing the LR-space. This finer and more thorough exploration of the possible design choices results in improved accuracy as well as reduction in parameters, FLOPS, and latency of a CNN model. Results demonstrate that the SVD-NAS achieves 2.06-12.85pp higher accuracy on ImageNet than state-of-the-art methods under the data-limited problem settings. SVD-NAS is open-sourced at http s://github.com/Yu-Zhewen/SVD-NAS.

**********

Multi-Level Contrastive Learning for Self-Supervised Vision Transformers

Shentong Mo, Zhun Sun, Chao Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2778-2787

Recent studies aim to establish contrastive self-supervised learning (CSL) algor ithms specialized for the family of Vision Transformers (ViTs) to make them func tion normally as ordinary convolutional-based backbones in the training progress . Despite obtaining promising performance on related downstream tasks, one compe lling property of the ViTs is ignored in those approaches. As previous studies h ave demonstrated, vision transformers benefit from the early stage global attent ion mechanics, obtaining feature representations that contain information from d istant patches, even in their shallow layers. Motivated by this, we present a si mple yet effective framework to facilitate the self-supervised feature learning of transformer-based vision architectures, namely, Multi-level Contrastive learn ing for Vision Transformers (MCVT). Specifically, we equip the vision transforme rs with individual-based (InfoNCE) and prototypical-based (ProtoNCE) contrastive loss in different stages of the architecture to capture low-level invariance an d high-level invariance between views of samples, respectively. We conduct exten sive experiments to demonstrate the effectiveness of the proposed method, using two well-known vision transformer backbones, on several vision downstream tasks, including linear classification, detection, and semantic segmentation.

**********

Do Pre-Trained Models Benefit Equally in Continual Learning?

Kuan-Ying Lee, Yuanyi Zhong, Yu-Xiong Wang; Proceedings of the IEEE/CVF Winter C onference on Applications of Computer Vision (WACV), 2023, pp. 6485-6493

A large part of the continual learning (CL) literature focuses on developing alg orithms for models trained from scratch. While these algorithms work great with from-sc ratch trained models on widely used CL benchmarks, they show dramatic pe rformance drops on more complex datasets (e.g., Split-CUB200). Pre-trained model s, widely used to transfer knowledge to downstream tasks, could enhance these me thods to be applicable in more realistic scenarios. However, surprisingly, impro vements in CL algorithms from pre-training are inconsistent. For instance, while Incremental Classifier and Representation Learning (iCaRL) underperforms Superv ised Contrastive Replay (SCR) when trained from scratch, it outperforms SCR when both are initialized with a pre-trained model. This indicates the paradigm curr ent CL literature follows, where all methods are compared in from-scratch traini ng, is not well reflective of the true CL objective and desired progress. Furthe rmore, we found 1) CL algorithms that exert less regularization benefit more fro m a pre-trained model; 2) a model pre-trained with a larger dataset (WebImageTex t in Contrastive Language-Image Pre-training (CLIP) vs. ImageNet) does not guara ntee a better improvement. Based on these findings, we introduced a simple yet e ffective baseline that employs minimum regularization and leverages the more ben eficial pre-trained model, which outperforms state-of-the-art methods when pre-t raining is applied. Our code is available at https://github.com/eric11220/pretra ined-models-in-CL.

**********
Cross-Domain Video Anomaly Detection Without Target Domain Adaptation
Abhishek Aich, Kuan-Chuan Peng, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2579-2591
Most cross-domain unsupervised Video Anomaly Detection (VAD) works assume that at least few task-relevant target domain training data are available for adaptation from the source to the target domain. However, this requires laborious model-tuning by the end-user who may prefer to have a system that works "out-of-the-box". To address such practical scenarios, we identify a novel target domain (inference-time) VAD task where no target domain training data are available. To this end, we propose a new 'Zero-shot Cross-domain Video Anomaly Detection (zxvad)' framework that includes a future-frame prediction generative model setup. Different from prior future-frame prediction models, our model uses a novel Normalcy Classifier module to learn the features of normal event videos by learning how such features are different "relative" to features in pseudo-abnormal examples. A novel Untrained Convolutional Neural Network based Anomaly Synthesis module crafts these pseudo-abnormal examples by adding foreign objects in normal video frames with no extra training cost. With our novel relative normalcy feature learning strategy, zxvad generalizes and learns to distinguish between normal and abnormal frames in a new target domain without adaptation during inference. Through evaluations on common datasets, we show that zxvad outperforms the state-of-the-art (SOTA), regardless of whether task-relevant (i.e., VAD) source training data are available or not. Lastly, zxvad also beats the SOTA methods in inference-time efficiency metrics including the model size, total parameters, GPU energy consumption, and GMACs.
**********
TeST: Test-Time Self-Training Under Distribution Shift
Samarth Sinha, Peter Gehler, Francesco Locatello, Bernt Schiele; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2759-2769
Despite their recent success, deep neural networks continue to perform poorly when they encounter distribution shifts at test time. Many recently proposed approaches try to counter this by aligning the model to the new distribution prior to inference. With no labels available this requires unsupervised objectives to adapt the model on the observed test data. In this paper, we propose Test-Time Self-Training (TeST): a technique that takes as input a model trained on some source data and a novel data distribution at test time, and learns invariant and robust representations using a student-teacher framework. We find that models adapted using TeST significantly improve over baseline test-time adaptation algorithms. TeST achieves competitive performance to modern domain adaptation algorithms [4,43], while having access to 5-10x less data at time of adaption. We thoroughly evaluate a variety of baselines on two tasks: object detection and image segmentation and find that models adapted with TeST. We find that TeST sets the new state-of-the art for test-time domain adaptation algorithms.
**********
IDD-3D: Indian Driving Dataset for 3D Unstructured Road Scenes
Shubham Dokania, A. H. Abdul Hafez, Anbumani Subramanian, Manmohan Chandraker, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4482-4491
Autonomous driving and assistance systems rely on annotated data from traffic and road scenarios to model and learn the various object relations in complex real-world scenarios. Preparation and training of deploy-able deep learning architectures require the models to be suited to different traffic scenarios and adapt to different situations. Currently, existing datasets, while large-scale, lack such diversities and are geographically biased towards mainly developed cities. An unstructured and complex driving layout found in several developing countries such as India poses a challenge to these models due to the sheer degree of variations in the object types, densities, and locations. To facilitate better research toward accommodating such scenarios, we build a new dataset,  IDD-3D , which c

onsists of multi-modal data from multiple cameras and LiDAR sensors with 12k annotated driving LiDAR frames across various traffic scenarios. We discuss the need for this dataset through statistical comparisons with existing datasets and highlight benchmarks on standard 3D object detection and tracking tasks in complex layouts.
**********

CoNMix for Source-Free Single and Multi-Target Domain Adaptation
Vikash Kumar, Rohit Lal, Himanshu Patil, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4178-4188
This work introduces the novel task of Source-free Multi-target Domain Adaptation and proposes adaptation framework comprising of Consistency with Nuclear-Norm Maximization and MixUp knowledge distillation (CoNMix) as a solution to this problem. The main motive of this work is to solve for Single and Multi target Domain Adaptation (SMTDA) for the source-free paradigm, which enforces a constraint where the labeled source data is not available during target adaptation due to various privacy-related restrictions on data sharing. The source-free approach leverages target pseudo labels, which can be noisy, to improve the target adaptation. We introduce consistency between label preserving augmentations and utilize pseudo label refinement methods to reduce noisy pseudo labels. Further, we propose novel MixUp Knowledge Distillation (MKD) for better generalization on multiple target domains using various source-free STDA models. We also show that the Vision Transformer (VT) backbone gives better feature representation with improved domain transferability and class discriminability. Our proposed framework achieves the state-of-the-art (SOTA) results in various paradigms of source-free STDA and MTDA settings on popular domain adaptation datasets like Office-Home, Office-Caltech, and DomainNet. Project Page: https://sites.google.com/view/conmix-vcl
**********

Temporal Feature Enhancement Dilated Convolution Network for Weakly-Supervised Temporal Action Localization
Jianxiong Zhou, Ying Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6028-6037
Weakly-supervised Temporal Action Localization (WTAL) aims to classify and localize action instances in untrimmed videos with only video-level labels. Existing methods typically use snippet-level RGB and optical flow features extracted from pre-trained extractors directly. Because of two limitations: the short temporal span of snippets and the inappropriate initial features, these WTAL methods suffer from the lack of effective use of temporal information and have limited performance. In this paper, we propose the Temporal Feature Enhancement Dilated Convolution Network (TFE-DCN) to address these two limitations. The proposed TFE-DCN has an enlarged receptive field that covers a long temporal span to observe the full dynamics of action instances, which makes it powerful to capture temporal dependencies between snippets. Furthermore, we propose the Modality Enhancement Module that can enhance RGB features with the help of enhanced optical flow features, making the overall features appropriate for the WTAL task. Experiments conducted on THUMOS'14 and ActivityNet v1.3 datasets show that our proposed approach far outperforms state-of-the-art WTAL methods.
**********

Meta-Auxiliary Learning for Future Depth Prediction in Videos
Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, Jin Tang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5756-5765
We consider a new problem of future depth prediction in video. Given a sequence of observed frames, the goal is to predict the depth map of a future frame that has not been observed yet. Depth estimation plays a vital role for scene understanding and decision-making in intelligent systems. Predicting future depth maps can be valuable for autonomous vehicles to anticipate the behaviors of their surrounding objects. Our proposed model for this problem has a two-branch architecture. One branch is for the primary task of future depth estimation. The other branch is for an auxiliary task of image reconstruction. The auxiliary branch can

act as a regularization. Inspired by some recent work on test-time adaption, we use the auxiliary task during testing to adapt the model to a specific test video. We also propose a novel meta-auxiliary learning that learn the model specifically for the purpose of effective test-time adaptation. Experimental results demonstrate that our proposed approach significantly outperforms other alternative methods.

**********

Large-to-Small Image Resolution Asymmetry in Deep Metric Learning

Pavel Suma, Giorgos Tolias; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1451-1460

Deep metric learning for vision is trained by optimizing a representation network to map (non-)matching image pairs to (non-)similar representations. During testing, which typically corresponds to image retrieval, both database and query examples are processed by the same network to obtain the representation used for similarity estimation and ranking. In this work, we explore an asymmetric setup by light-weight processing of the query at a small image resolution to enable fast representation extraction. The goal is to obtain a network for database examples that is trained to operate on large resolution images and benefits from fine-grained image details, and a second network for query examples that operates on small resolution images but preserves a representation space aligned with that of the database network. We achieve this with a distillation approach that transfers knowledge from a fixed teacher network to a student via a loss that operates per image and solely relies on coupled augmentations without the use of any labels. In contrast to prior work that explores such asymmetry from the point of view of different network architectures, this work uses the same architecture but modifies the image resolution. We conclude that resolution asymmetry is a better way to optimize the performance/efficiency trade-off than architecture asymmetry. Evaluation is performed on three standard deep metric learning benchmarks, namely CUB200, Cars196, and SOP. Code: https://github.com/pavelsuma/raml

**********

SERF: Towards Better Training of Deep Neural Networks Using Log-Softplus ERror Activation Function

Sayan Nag, Mayukh Bhattacharyya, Anuraag Mukherjee, Rohit Kundu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5324-5333

Activation functions play a pivotal role in determining the training dynamics and neural network performance. The widely adopted activation function ReLU despite being simple and effective has few disadvantages including the Dying ReLU problem. In order to tackle such problems, we propose a novel activation function called Serf which is self-regularized and nonmonotonic in nature. Like Mish, Serf also belongs to the Swish family of functions. Based on several experiments on computer vision (image classification and object detection) and natural language processing (machine translation, sentiment classification and multimodal entailment) tasks with different state-of-the-art architectures, it is observed that Serf vastly outperforms ReLU (baseline) and other activation functions including both Swish and Mish, with a markedly bigger margin on deeper architectures. Ablation studies further demonstrate that Serf based architectures perform better than those of Swish and Mish in varying scenarios, validating the effectiveness and compatibility of Serf with varying depth, complexity, optimizers, learning rates, batch sizes, initializers and dropout rates. Finally, we investigate the mathematical relation between Swish and Serf, thereby showing the impact of preconditioner function ingrained in the first derivative of Serf which provides a regularization effect making gradients smoother and optimization faster.

**********

AVE-CLIP: AudioCLIP-Based Multi-Window Temporal Transformer for Audio Visual Event Localization

Tanvir Mahmud, Diana Marculescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5158-5167

An audio-visual event (AVE) is denoted by the correspondence of the visual and auditory signals in a video segment. Precise localization of the AVEs is very cha

llenging since it demands effective multi-modal feature correspondence to ground the short and long range temporal interactions. Existing approaches struggle in capturing the different scales of multi-modal interaction due to ineffective multi-modal training strategies. To overcome this limitation, we introduce AVE-CLIP, a novel framework that integrates the AudioCLIP pre-trained on large-scale audio-visual data with a multi-window temporal transformer to effectively operate on different temporal scales of video frames. Our contributions are three-fold: (1) We introduce a multi-stage training framework to incorporate AudioCLIP pre-trained with audio-image pairs into the AVE localization task on video frames through contrastive fine-tuning, effective mean video feature extraction, and multi-scale training phases. (2) We propose a multi-domain attention mechanism that operates on both temporal and feature domains over varying timescales to fuse the local and global feature variations. (3) We introduce a temporal refining scheme with event-guided attention followed by a simple-yet-effective post processing step to handle significant variations of the background over diverse events. Our method achieves state-of-the-art performance on the publicly available AVE dataset with 5.9% mean accuracy improvement which proves its superiority over existing approaches.

**********

Misclassifications of Contact Lens Iris PAD Algorithms: Is It Gender Bias or Environmental Conditions?

Akshay Agarwal, Nalini Ratha, Afzel Noore, Richa Singh, Mayank Vatsa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 961-970

One of the critical steps in biometrics pipeline is detection of presentation attacks, a physical adversary. Several presentation (adversary) attack detection (PAD) algorithms, including iris PAD, are proposed and have shown superlative performance. However, a recent study, on a small-scale database, has highlighted that iris PAD may have gender biases. In this research, we present a rigorous study on gender bias in iris presentation attack detection algorithms using a large-scale and gender-balanced database. The paper provides several interesting observations which can help in building future presentation attack detection algorithms with aim of fair treatment of each demography. In addition, we also present a robust iris presentation attack detection algorithm by combining gender-covariate biased classifiers. The proposed robust classifier not only reduces the difference in accuracy between different genders but also improves the overall performance of the PAD system.

**********

Empirical Generalization Study: Unsupervised Domain Adaptation vs. Domain Generalization Methods for Semantic Segmentation in the Wild

Fabrizio J. Piva, Daan de Geus, Gijs Dubbelman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 499-508

For autonomous vehicles and mobile robots to safely operate in the real world, i.e., the wild, scene understanding models should perform well in the many different scenarios that can be encountered. In reality, these scenarios are not all represented in the model's training data, leading to poor performance. To tackle this, current training strategies attempt to either exploit additional unlabeled data with unsupervised domain adaptation (UDA), or to reduce overfitting using the limited available labeled data with domain generalization (DG). However, it is not clear from current literature which of these methods allows for better generalization to unseen data from the wild. Therefore, in this work, we present an evaluation framework in which the generalization capabilities of state-of-the-art UDA and DG methods can be compared fairly. From this evaluation, we find that UDA methods, which leverage unlabeled data, outperform DG methods in terms of generalization, and can deliver similar performance on unseen data as fully-supervised training methods that require all data to be labeled. We show that semantic segmentation performance can be increased up to 30% for a priori unknown data without using any extra labeled data.

**********

VLC-BERT: Visual Question Answering With Contextualized Commonsense Knowledge

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, Vered Shwartz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1155-1165

There has been a growing interest in solving Visual Question Answering (VQA) tasks that require the model to reason beyond the content present in the image. In this work, we focus on questions that require commonsense reasoning. In contrast to previous methods which inject knowledge from static knowledge bases, we investigate the incorporation of contextualized knowledge using Commonsense Transformer (COMET), an existing knowledge model trained on human-curated knowledge bases. We propose a method to generate, select, and encode external commonsense knowledge alongside visual and textual cues in a new pre-trained Vision-Language-Commonsense transformer model, VLC-BERT. Through our evaluation on the knowledge-intensive OK-VQA and A-OKVQA datasets, we show that VLC-BERT is capable of outperforming existing models that utilize static knowledge bases. Furthermore, through a detailed analysis, we explain which questions benefit, and which don't, from contextualized commonsense knowledge from COMET.

**********

Feature Disentanglement Learning With Switching and Aggregation for Video-Based Person Re-Identification

Minjung Kim, MyeongAh Cho, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1603-1612

In video person re-identification (Re-ID), the network must consistently extract features of the target person from successive frames. Existing methods tend to focus only on how to use temporal information, which often leads to networks being fooled by similar appearances and same backgrounds. In this paper, we propose a Disentanglement and Switching and Aggregation Network (DSANet), which segregates the features representing identity and features based on camera characteristics, and pays more attention to ID information. We also introduce an auxiliary task that utilizes a new pair of features created through switching and aggregation to increase the network's capability for various camera scenarios. Furthermore, we devise a Target Localization Module (TLM) that extracts robust features against a change in the position of the target according to the frame flow and a Frame Weight Generation (FWG) that reflects temporal information in the final representation. Various loss functions for disentanglement learning are designed so that each component of the network can cooperate while satisfactorily performing its own role. Quantitative and qualitative results from extensive experiments demonstrate the superiority of DSANet over state-of-the-art methods on three benchmark datasets.

**********

OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping

Junshi Xia, Naoto Yokoya, Bruno Adriano, Clifford Broni-Bediako; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6254-6264

We introduce OpenEarthMap, a benchmark dataset, for global high-resolution land cover mapping. OpenEarthMap consists of 2.2 million segments of 5000 aerial and satellite images covering 97 regions from 44 countries across 6 continents, with manually annotated 8-class land cover labels at a 0.25--0.5m ground sampling distance. Semantic segmentation models trained on the OpenEarthMap generalize worldwide and can be used as off-the-shelf models in a variety of applications. We evaluate the performance of state-of-the-art methods for unsupervised domain adaptation and present challenging problem settings suitable for further technical development. We also investigate lightweight models using automated neural architecture search for limited computational resources and fast mapping. The dataset will be made publicly available.

**********

Semantics Guided Contrastive Learning of Transformers for Zero-Shot Temporal Activity Detection

Sayak Nag, Orpaz Goldstein, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6243-6253

Zero-shot temporal activity detection (ZSTAD) is the problem of simultaneous tem

poral localization and classification of activity segments that are previously u
nseen during training. This is achieved by transferring the knowledge learned fr
om semantically-related seen activities. This ability to reason about unseen con
cepts without supervision makes ZSTAD very promising for applications where the
acquisition of annotated training videos is difficult. In this paper, we design
a transformer-based framework titled TranZAD, which streamlines the detection of
 unseen activities by casting ZSTAD as a direct set-prediction problem, removing
 the need for hand-crafted designs and manual post-processing. We show how a sem
antic information-guided contrastive learning strategy can effectively train Tra
nZAD for the zero-shot setting, enabling the efficient transfer of knowledge fro
m the seen to the unseen activities. To reduce confusion between unseen activiti
es and unrelated background information in videos, we introduce a more efficient
 method of computing the background class embedding by dynamically adapting it a
s part of the end-to-end learning. Additionally, unlike existing work on ZSTAD,
we do not assume the knowledge of which classes are unseen during training and u
se the visual and semantic information of only the seen classes for the knowledg
e transfer. This makes TranZAD more viable for practical scenarios, which we eva
luate by conducting extensive experiments on Thumos'14 and Charades.
**********

Real-Time Concealed Weapon Detection on 3D Radar Images for Walk-Through Screeni
ng System

Nagma S. Khan, Kazumine Ogura, Eric Cosatto, Masayuki Ariyoshi; Proceedings of t
he IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, p
p. 673-681

This paper presents a framework for real-time concealed weapon detection (CWD) o
n 3D radar images for walk-through screening systems. The walk-through screening
 system aims to ensure security in crowded areas by performing CWD on walking pe
rsons, hence it requires an accurate and real-time detection approach. To ensure
 accuracy, a weapon needs to be detected irrespective of its 3D orientation, thu
s we use the 3D radar images as detection input. For achieving real-time, we ref
ormulate classic U-Net based segmentation networks to perform 3D detection tasks
. Our 3D segmentation network predicts peak-shaped probability map, instead of v
oxel-wise masks, to enable position inference by elementary peak detection opera
tion on the predicted map. In the peak-shaped probability map, the peak marks th
e weapon's position. So, weapon detection task translates to peak detection on t
he probability map. A Gaussian function is used to model weapons in the probabil
ity map. We experimentally validate our approach on realistic 3D radar images ob
tained from a walk-through weapon screening system prototype. Extensive ablation
 studies verify the effectiveness of our proposed approach over existing convent
ional approaches. The experimental results demonstrate that our proposed approac
h can perform accurate and real-time CWD, thus making it suitable for practical
applications of walk-through screening.
**********

AFPSNet: Multi-Class Part Parsing Based on Scaled Attention and Feature Fusion

Njuod Alsudays, Jing Wu, Yu-Kun Lai, Ze Ji; Proceedings of the IEEE/CVF Winter C
onference on Applications of Computer Vision (WACV), 2023, pp. 4033-4042

Multi-class part parsing is a dense prediction task that seeks to simultaneously
 detect multiple objects and the semantic parts within these objects in the scen
e. This problem is important in providing detailed object understanding, but is
challenging due to the existence of both class-level and part-level ambiguities.
 In this paper, we propose to integrate an attention refinement module and a fea
ture fusion module to tackle the part-level ambiguity. The attention refinement
module aims to enhance the feature representations by focusing on important feat
ures. The feature fusion module aims to improve the fusion operation for differe
nt scales of features. We also propose an object-to-part training strategy to ta
ckle the class-level ambiguity, which improves the localization of parts by expl
oiting prior knowledge of objects. The experimental results demonstrated the eff
ectiveness of the proposed modules and the training strategy, and showed that ou
r proposed method achieved state-of-the-art performance on the benchmark dataset
.

**********
The Change You Want To See
Ragav Sachdeva, Andrew Zisserman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3993-4002
We live in a dynamic world where things change all the time. Given two images of the same scene, being able to automatically detect the changes in them has practical applications in a variety of domains. In this paper, we tackle the change detection problem with the goal of detecting "object-level" changes in an image pair despite differences in their viewpoint and illumination. To this end, we make the following four contributions: (i) we propose a scalable methodology for obtaining a large-scale change detection training dataset by leveraging existing object segmentation benchmarks; (ii) we introduce a co-attention based novel architecture that is able to implicitly determine correspondences between an image pair and find changes in the form of bounding box predictions; (iii) we contribute four evaluation datasets that cover a variety of domains and transformations, including synthetic image changes, real surveillance images of a 3D scene, and synthetic 3D scenes with camera motion; (iv) we evaluate our model on these four datasets and demonstrate zero-shot and beyond training transformation generalization. The code, datasets and pre-trained model can be found at our project page: https://www.robots.ox.ac.uk/ vgg/research/cyws/
**********
Kernel-Aware Burst Blind Super-Resolution
Wenyi Lian, Shanglian Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4892-4902
Burst super-resolution technique provides a possibility of restoring rich details from low-quality images. However, since real world low-resolution (LR) images in practical applications have multiple complicated and unknown degradations, existing non-blind (e.g., bicubic) designed networks usually suffer severe performance drop in recovering high-resolution (HR) images. In this paper, we address the problem of reconstructing HR images from raw burst sequences acquired from modern handheld devices. The central idea is a kernel-guided strategy which can solve the burst SR problem with two steps: kernel estimation and HR image restoration. The former estimates burst kernels from raw inputs, while the latter predicts the super-resolved image based on the estimated kernels. Furthermore, we introduce a pyramid kernel-aware deformable alignment module which can effectively align the raw images with consideration of the blurry priors. Extensive experiments on synthetic and real-world datasets demonstrate that the proposed method can perform favorable state-of-the-art performance in the burst SR problem.
**********
Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection
Hanqiu Deng, Zhaoxiang Zhang, Shihao Zou, Xingyu Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2634-2643
Anomaly detection in video surveillance aims to detect anomalous frames whose properties significantly differ from normal patterns. Anomalies in videos can occur in both spatial appearance and temporal motion, making unsupervised video anomaly detection challenging. To tackle this problem, we investigate forward and backward motion continuity between adjacent frames and propose a new video anomaly detection paradigm based on bi-directional frame interpolation. The proposed framework consists of an optical flow estimation network and an interpolation network jointly optimized end-to-end to synthesize a middle frame from its nearest two frames. We further introduce a novel dynamic memory mechanism to balance memory sparsity and normality representation diversity, which attenuates abnormal features in frame interpolation without affecting normal prototypes. In inference, interpolation error and dynamic memory error are fused as anomaly scores. The proposed bi-directional interpolation design improves normal frame synthesis, lowering the false alarm rate of anomaly appearance; meanwhile, the implicit "regular" motion constraint in our optical flow estimation and the novel dynamic memory mechanism play blocking roles in interpolating abnormal frames, increasing the system's sensitivity to anomalies. Extensive experiments on public benchmarks d

emonstrate the superiority of the proposed framework over prior arts.
**********
Is Your Noise Correction Noisy? PLS: Robustness To Label Noise With Two Stage Detection
Paul Albert, Eric Arazo, Tarun Krishna, Noel E. O'Connor, Kevin McGuinness; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 118-127
Designing robust algorithms capable of training accurate neural networks on uncurated datasets from the web has been the subject of much research as it reduces the need for time consuming human labor. The focus of many previous research contributions has been on the detection of different types of label noise; however, this paper proposes to improve the correction accuracy of noisy samples once they have been detected. In many state-of-the-art contributions, a two phase approach is adopted where the noisy samples are detected before guessing a corrected pseudo-label in a semi-supervised fashion. The guessed pseudo-labels are then used in the supervised objective without ensuring that the label guess is likely to be correct. This can lead to confirmation bias, which reduces the noise robustness. Here we propose the pseudo-loss, a simple metric that we find to be strongly correlated with pseudo-label correctness on noisy samples. Using the pseudo-loss, we dynamically down weight under-confident pseudo-labels throughout training to avoid confirmation bias and improve the network accuracy. We additionally propose to use a confidence guided contrastive objective that learns robust representation on an interpolated objective between class bound (supervised) for confidently corrected samples and unsupervised representation for under-confident label corrections. Experiments demonstrate the state-of-the-art performance of our Pseudo-Loss Selection (PLS) algorithm on a variety of benchmark datasets including curated data synthetically corrupted with in-distribution and out-of-distribution noise, and two real world web noise datasets. Our experiments are fully reproducible github.com/PaulAlbert31/PLS.
**********
PP4AV: A Benchmarking Dataset for Privacy-Preserving Autonomous Driving
Linh Trinh, Phuong Pham, Hoang Trinh, Nguyen Bach, Dung Nguyen, Giang Nguyen, Huy Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1206-1215
Massive data collected on public roads for autonomous driving has become more popular in many locations in the world. More collected data leads to more concerns about data privacy, including but not limited to pedestrian faces and surrounding vehicle license plates, which urges for robust solutions for detecting and anonymizing them in realistic road-driving scenarios. Existing public datasets for both face and license plate detection are either not focused on autonomous driving or only in parking lots. In this paper, we introduce a challenging public dataset for face and license plate detection in autonomous driving domain. The dataset is aggregated from visual data that is available in public domain, to cover scenarios from six European cities, including daytime and nighttime, annotated with both faces and license plates. All of the images feature a variety of poses and sizes for both faces and license plates. Our dataset offers not only a benchmark for evaluating data anonymization models but also data to get more insights about privacy-preserving autonomous driving. The experimental results showed that 1) current generic state-of-the-art face and/or license plate detection models do not perform well on a realistic and diverse road-driving dataset like ours, 2) our model trained with autonomous driving data (even with soft-labeling data) outperformed strong but generic models, and 3) the size of faces and license plates is an important factor for evaluating and optimizing the performance of privacy-preserving autonomous driving. The annotation of dataset as well as baseline model and results are available at our github: https://github.com/khaclinh/pp4av.
**********
Heightfields for Efficient Scene Reconstruction for AR
Jamie Watson, Sara Vicente, Oisin Mac Aodha, Clément Godard, Gabriel Brostow, Michael Firman; Proceedings of the IEEE/CVF Winter Conference on Applications of C

omputer Vision (WACV), 2023, pp. 5850-5860
3D scene reconstruction from a sequence of posed RGB images is a cornerstone tas k for computer vision and augmented reality (AR). While depth-based fusion is th e foundation of most real-time approaches for 3D reconstruction, recent learning based methods that operate directly on RGB images can achieve higher quality re constructions, but at the cost of increased runtime and memory requirements, mak ing them unsuitable for AR applications. We propose an efficient learning-based method that refines the 3D reconstruction obtained by a traditional fusion appro ach. By leveraging a top-down heightfield representation, our method remains rea l-time while approaching the quality of other learning-based methods. Despite be ing a simplification, our heightfield is perfectly appropriate for robotic path planning or augmented reality character placement. We outline several innovation s that push the performance beyond existing top-down prediction baselines, and w e present an evaluation framework on the challenging ScanNetV2 dataset, targetin g AR tasks. Ultimately, we show that our method improves over the baselines for AR applications. Full code and pretrained models will be released on acceptance.
**********

MFFN: Multi-View Feature Fusion Network for Camouflaged Object Detection
Dehua Zheng, Xiaochen Zheng, Laurence T. Yang, Yuan Gao, Chenlu Zhu, Yiheng Ruan ; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi on (WACV), 2023, pp. 6232-6242
Recent research about camouflaged object detection (COD) aims to segment highly concealed objects hidden in complex surroundings. The tiny, fuzzy camouflaged ob jects result in visually indistinguishable properties. However, current single-v iew COD detectors are sensitive to background distractors. Therefore, blurred bo undaries and variable shapes of the camouflaged objects are challenging to be fu lly captured with a single-view detector. To overcome these obstacles, we propos e a behavior-inspired framework, called Multi-view Feature Fusion Network (MFFN) , which mimics the human behaviors of finding indistinct objects in images, i.e. , observing from multiple angles, distances, perspectives. Specifically, the key idea behind it is to generate multiple ways of observation (multi-view) by data augmentation and apply them as inputs. MFFN captures critical boundary and sema ntic information by comparing and fusing extracted multi-view features. In addit ion, our MFFN exploits the dependence and interaction between views and channels . Specifically, our methods leverage the complementary information between diffe rent views through a two-stage attention module called Co-attention of Multi-vie w (CAMV). And we design a local-overall module called Channel Fusion Unit (CFU) to explore the channel-wise contextual clues of diverse feature maps in an itera tive manner. The experiment results show that our method performs favorably agai nst existing state-of-the-art methods via training with the same data. The code will be available at https://github.com/dwardzheng/MFFN_COD.
**********

Federated Learning for Commercial Image Sources
Shreyansh Jain, Koteswar Rao Jerripothula; Proceedings of the IEEE/CVF Winter Co nference on Applications of Computer Vision (WACV), 2023, pp. 6534-6543
Federated Learning is a collaborative machine learning paradigm that enables mul tiple clients to learn a global model without exposing their data to each other. Consequently, it provides a secure learning platform with privacy-preserving ca pabilities. This paper introduces a new dataset containing 23,326 images collect ed from eight different commercial sources and classified into 31 categories, si milar to the Office-31 dataset. To the best of our knowledge, this is the first image classification dataset specifically designed for Federated Learning. We al so propose two new Federated Learning algorithms, namely Fed-Cyclic and Fed-Star . In Fed-Cyclic, a client receives weights from its previous client, updates the m through local training, and passes them to the next client, thus forming a cyc lic topology. In Fed-Star, a client receives weights from all other clients, upd ates its local weights through pre-aggregation (to address statistical heterogen eity) and local training, and sends its updated local weights to all other clien ts, thus forming a star-like topology. Our experiments reveal that both algorith ms perform better than existing baselines on our newly introduced dataset.

**********
Adaptive Local-Component-Aware Graph Convolutional Network for One-Shot Skeleton-Based Action Recognition
Anqi Zhu, Qiuhong Ke, Mingming Gong, James Bailey; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6038-6047
Skeleton-based action recognition receives increasing attention because skeleton sequences reduce training complexity by eliminating visual information irrelevant to actions. To further improve sample efficiency, meta-learning-based one-shot learning solutions were developed for skeleton-based action recognition. These methods predict by finding the nearest neighbors according to the similarity between instance-level global embedding. However, such measurement holds unstable representativity due to inadequate generalized learning on the averaged local invariant and noisy features, while intuitively, steady and fine-grained recognition relies on determining key local body movements. To address this limitation, we present the Adaptive Local-Component-aware Graph Convolutional Network, which replaces the comparison metric with a focused sum of similarity measurements on aligned local embedding of action-critical spatial/temporal segments. Comprehensive one-shot experiments on the public benchmark of NTU-RGB+D 120 indicate that our method provides a stronger representation than the global embedding and helps our model reach state-of-the-art.
**********
Inducing Data Amplification Using Auxiliary Datasets in Adversarial Training
Saehyung Lee, Hyungyu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4551-4560
Several recent studies have shown that the use of extra in-distribution data can lead to a high level of adversarial robustness. However, there is no guarantee that it will always be possible to obtain sufficient extra data for a selected dataset. In this paper, we propose a biased multi-domain adversarial training (BiaMAT) method that induces training data amplification on a primary dataset using publicly available auxiliary datasets, without requiring the class distribution match between the primary and auxiliary datasets. The proposed method can achieve increased adversarial robustness on a primary dataset by leveraging auxiliary datasets via multi-domain learning. Specifically, data amplification on both robust and non-robust features can be accomplished through the application of BiaMAT as demonstrated through a theoretical and empirical analysis. Moreover, we demonstrate that while existing methods are vulnerable to negative transfer due to the distributional discrepancy between auxiliary and primary data, the proposed method enables neural networks to flexibly leverage diverse image datasets for adversarial training by successfully handling the domain discrepancy through the application of a confidence-based selection strategy. The code and pre-trained models of our study are available at: https://github.com/BiaMAT/BiaMAT_under_review.
**********
AttTrack: Online Deep Attention Transfer for Multi-Object Tracking
Keivan Nalaie, Rong Zheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1654-1663
Multi-object tracking (MOT) is a vital component of intelligent video analytics applications such as surveillance and autonomous driving. The time and storage complexity required to execute deep learning models for visual object tracking hinder their adoption on embedded devices with limited computing power. In this paper, we aim to accelerate MOT by transferring the knowledge from high-level features of a complex network (teacher) to a lightweight network (student) at both training and inference times. The proposed AttTrack framework has three key components: 1) cross-model feature learning to align intermediate representations from the teacher and student models, 2) interleaving the execution of the two models at inference time, and 3) incorporating the updated predictions from the teacher model as prior knowledge to assist the student model. Experiments on pedestrian tracking tasks are conducted on the MOT17 and MOT15 datasets using two different object detection backbones YOLOv5 and DLA34 show that AttTrack can significantly improve student model tracking performance while sacrificing only minor deg

radation of tracking speed.
**********

Pruning-Guided Curriculum Learning for Semi-Supervised Semantic Segmentation
Heejo Kong, Gun-Hee Lee, Suneung Kim, Seong-Whan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5914-5923
This study focuses on improving the quality of pseudo-labeling in the context of semi-supervised semantic segmentation. Previous studies have adopted confidence thresholding to reduce erroneous predictions in pseudo-labeled data and to enhance their qualities. However, numerous pseudo-labels with high confidence scores exist in the early training stages even though their predictions are incorrect, and this ambiguity limits confidence thresholding substantially. In this paper, we present a novel method to resolve the ambiguity of confidence scores with the guidance of network pruning. A recent finding showed that network pruning severely impairs the network generalization ability on samples that are not yet well learned or represented. Inspired by this finding, we refine the confidence scores by reflecting the extent to which the predictions are affected by pruning. Furthermore, we adopted a curriculum learning strategy for the confidence score, which enables the network to learn gradually from easy to hard samples. This approach resolves the ambiguity by suppressing the learning of noisy pseudo-labels, the confidence scores of which are difficult to trust owing to insufficient training in the early stages. Extensive experiments on various benchmarks demonstrate the superiority of our framework over state-of-the-art alternatives.
**********

ML-Decoder: Scalable and Versatile Classification Head
Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, Asaf Noy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 32-41
In this paper, we introduce ML-Decoder, a new attention-based classification head. ML-Decoder predicts the existence of class labels via queries, and enables better utilization of spatial data compared to global average pooling. By redesigning the decoder architecture, and using a novel group-decoding scheme, ML-Decoder is highly efficient, and can scale well to thousands of classes. Compared to using a larger backbone, ML-Decoder consistently provides a better speed-accuracy trade-off. ML-Decoder is also versatile - it can be used as a drop-in replacement for various classification heads, and generalize to unseen classes when operated with word queries. Novel query augmentations further improve its generalization ability. Using ML-Decoder, we achieve state-of-the-art results on several classification tasks: on MS-COCO multi-label, we reach 91.1% mAP; on NUS-WIDE zero-shot, we reach 31.1% ZSL mAP; and on ImageNet single-label, we reach with vanilla ResNet50 backbone a new top score of 80.7%, without extra data or distillation. Public code will be available.
**********

Zero-Shot Versus Many-Shot: Unsupervised Texture Anomaly Detection
Toshimichi Aota, Lloyd Teh Tzer Tong, Takayuki Okatani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5564-5572
Research on unsupervised anomaly detection (AD) has recently progressed, significantly increasing detection accuracy. This paper focuses on texture images and considers how few normal samples are needed for accurate AD. We first highlight the critical nature of the problem that previous studies have overlooked: accurate detection gets harder for anisotropic textures when image orientations are not aligned between inputs and normal samples. We then propose a zero-shot method, which detects anomalies without using a normal sample. The method is free from the issue of unaligned orientation between input and normal images. It assumes the input texture to be homogeneous, detecting image regions that break the homogeneity as anomalies. We present a quantitative criterion to judge whether this assumption holds for an input texture. Experimental results show the broad applicability of the proposed zero-shot method and its good performance comparable to or even higher than the state-of-the-art methods using hundreds of normal samples

. The code and data are available from https://drive.google.com/drive/folders/10OyPzvI3H6llCZBxKxFlKWt1Pw1tkMK1.
**********

Unsupervised Audio-Visual Lecture Segmentation
Darshan Singh S., Anchit Gupta, C. V. Jawahar, Makarand Tapaswi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5232-5241
Over the last decade, online lecture videos have become increasingly popular and have experienced a meteoric rise during the pandemic. However, video-language research has primarily focused on instructional videos or movies, and tools to help students navigate the growing online lectures are lacking. Our first contribution is to facilitate research in the educational domain, by introducing AVLectures, a large-scale dataset consisting of 86 courses with over 2,350 lectures covering various STEM subjects. Each course contains video lectures, transcripts, OCR outputs for lecture frames, and optionally lecture notes, slides, assignments, and related educational content that can inspire a variety of tasks. Our second contribution is introducing video lecture segmentation that splits lectures into bite-sized topics that show promise in improving learner engagement. We formulate lecture segmentation as an unsupervised task that leverages visual, textual, and OCR cues from the lecture, while clip representations are fine-tuned on a pretext self-supervised task of matching the narration with the temporally aligned visual content. We use these representations to generate segments using a temporally consistent 1-nearest neighbor algorithm, TW-FINCH. We evaluate our method on 15 courses and compare it against various visual and textual baselines, outperforming all of them. Our comprehensive ablation studies also identify the key factors driving the success of our approach.
**********

Diffeomorphic Image Registration With Neural Velocity Field
Kun Han, Shanlin Sun, Xiangyi Yan, Chenyu You, Hao Tang, Junayed Naushad, Haoyu Ma, Deying Kong, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1869-1879
Diffeomorphic image registration, offering smooth transformation and topology preservation, is required in many medical image analysis tasks.Traditional methods impose certain modeling constraints on the space of admissible transformations and use optimization to find the optimal transformation between two images. Specifying the right space of admissible transformations is challenging: the registration quality can be poor if the space is too restrictive, while the optimization can be hard to solve if the space is too general. Recent learning-based methods, utilizing deep neural networks to learn the transformation directly, achieve fast inference, but face challenges in accuracy due to the difficulties in capturing the small local deformations and generalization ability. Here we propose a new optimization-based method named DNVF (Diffeomorphic Image Registration with Neural Velocity Field) which utilizes deep neural network to model the space of admissible transformations. A multilayer perceptron (MLP) with sinusoidal activation function is used to represent the continuous velocity field and assigns a velocity vector to every point in space, providing the flexibility of modeling complex deformations as well as the convenience of optimization. Moreover, we propose a cascaded image registration framework (Cas-DNVF) by combining the benefits of both optimization and learning based methods, where a fully convolutional neural network (FCN) is trained to predict the initial deformation, followed by DNVF for further refinement. Experiments on two large-scale 3D MR brain scan datasets demonstrate that our proposed methods significantly outperform the state-of-the-art registration methods.
**********

Dense but Efficient VideoQA for Intricate Compositional Reasoning
Jihyeon Lee, Wooyoung Kang, Eun-Sol Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1114-1123
It is well known that most of the conventional video question answering (VideoQA) datasets consist of easy questions requiring simple reasoning processes. However, long videos inevitably contain complex and compositional semantic structures

along with the spatio-temporal axis, which requires a model to understand the compositional structures inherent in the videos. In this paper, we suggest a new compositional VideoQA method based on transformer architecture with a deformable attention mechanism to address the complex VideoQA tasks. The deformable attentions are introduced to sample a subset of informative visual features from the dense visual feature map to cover a temporally long range of frames efficiently. Furthermore, the dependency structure within the complex question sentences is also combined with the language embeddings to readily understand the relations among question words. Extensive experiments and ablation studies show that the suggested dense but efficient model outperforms other baselines.
**********

Multi-View Photometric Stereo Revisited
Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3126-3135
Multi-view photometric stereo (MVPS) is a preferred method for detailed and precise 3D acquisition of an object from images. Although popular methods for MVPS can provide outstanding results, they are often complex to execute and limited to isotropic material objects. To address such limitations, we present a simple, practical approach to MVPS, which works well for isotropic as well as other object material types such as anisotropic and glossy. The proposed approach in this paper exploits the benefit of uncertainty modeling in a deep neural network for a reliable fusion of photometric stereo (PS) and multi-view stereo (MVS) network predictions. Yet, contrary to the recently proposed state-of-the-art, we introduce neural volume rendering methodology for a trustworthy fusion of MVS and PS measurements. The advantage of introducing neural volume rendering is that it helps in the reliable modeling of objects with diverse material types, where existing MVS methods, PS methods, or both may fail. Furthermore, it allows us to work on neural 3D shape representation, which has recently shown outstanding results for many geometric processing tasks. Our suggested new loss function aims to fit the zero level set of the implicit neural function using the most certain MVS and PS network predictions coupled with weighted neural volume rendering cost. The proposed approach shows state-of-the-art results when tested extensively on several benchmark datasets.
**********

K-VQG: Knowledge-Aware Visual Question Generation for Common-Sense Acquisition
Kohei Uehara, Tatsuya Harada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4401-4409
Visual Question Generation (VQG) is a task to generate questions from images. When humans ask questions about an image, their goal is often to acquire some new knowledge. However, existing studies on VQG have mainly addressed question generation from answers or question categories, overlooking the objectives of knowledge acquisition. To introduce a knowledge acquisition perspective into VQG, we constructed a novel knowledge-aware VQG dataset called K-VQG. This is the first large, humanly annotated dataset in which questions regarding images are tied to structured knowledge. We also developed a new VQG model that can encode and use knowledge as the target for a question. The experiment results show that our model outperforms existing models on the K-VQG dataset. Our dataset is publicly available at https://uehara-mech.github.io/kvqg.
**********

VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset With Consistent Relationship Labels
Yue Qiu, Yoshiki Nagasaki, Kensho Hara, Hirokatsu Kataoka, Ryota Suzuki, Kenji Iwata, Yutaka Satoh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3351-3360
Spatio-temporal scene graph generation is an essential task in household activity recognition that aims to identify human-object interactions. Constructing a dataset with per-frame object region and consistent relationship annotations requires extremely high labor costs. Existing datasets sparsely annotate frames sampled from videos, resulting in the lack of dense spatio-temporal correlation in vi

deos. Additionally, existing datasets contain inconsistent relationship annotations, leading to the problem of learning ambiguous temporal associations. Moreover, existing datasets mainly discuss relationships that can be inferred from a single frame, ignoring the significance of temporal associations. To resolve those issues, we created a simulated dataset with per-frame consistent annotations and introduced a range of relationships requiring both spatial and temporal context. Most existing methods explore spatial correlations within single images and do not explicitly consider the dynamic changes across frames. Therefore, we proposed a tracking-based approach that explicitly grasps spatio-temporal human-object interactions while simultaneously localizing humans and objects. Our proposed approach achieved state-of-the-art performance on scene graph generation and outperformed existing methods in scene graph localization by large margins on the proposed dataset. Moreover, the experiments show the efficacy of pre-training on the proposed dataset while adapting to a previous benchmark consisting of real daily videos, indicating the potential of the proposed dataset in real-world scenarios.
**********

Towards Interpretable Video Anomaly Detection
Keval Doshi, Yasin Yilmaz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2655-2664
Most video anomaly detection approaches are based on data-intensive end-to-end trained neural networks, which extract spatiotemporal features from videos. The extracted feature representations in such approaches are not interpretable, which prevents the automatic identification of anomaly cause. To this end, we propose a novel framework which can explain the detected anomalous event in a surveillance video. In addition to monitoring objects independently, we also monitor the interactions between them to detect anomalous events and explain their root causes. Specifically, we demonstrate that the scene graphs obtained by monitoring the object interactions provide an interpretation for the context of the anomaly while performing competitively with respect to the recent state-of-the-art approaches. Moreover, the proposed interpretable method enables cross-domain adaptability (i.e., transfer learning in another surveillance scene), which is not feasible for most existing end-to-end methods due to the lack of sufficient labeled training data for every surveillance scene. The quick and reliable detection performance of the proposed method is evaluated both theoretically (through an asymptotic optimality proof) and empirically on the popular benchmark datasets.
**********

Weakly Supervised Cell-Instance Segmentation With Two Types of Weak Labels by Single Instance Pasting
Kazuya Nishimura, Ryoma Bise; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3185-3194
Cell instance segmentation that recognizes each cell boundary is an important task in cell image analysis. While deep learning-based methods have shown promising performances with a certain amount of training data, most of them require full annotations that show the boundary of each cell. Generating the annotation for cell segmentation is time-consuming and human labor. To reduce the annotation cost, we propose a weakly supervised segmentation method using two types of weak labels (one for cell type and one for nuclei position). Unlike general images, these two labels are easily obtained in phase-contrast images. The intercellular boundary, which is necessary for cell instance segmentation, cannot be directly obtained from these two weak labels, so to generate the boundary information, we propose a single instance pasting based on the copy-and-paste technique. First, we locate single-cell regions by counting cells and store them in a pool. Then, we generate the intercellular boundary by pasting the stored single-cell regions to the original image. Finally, we train a boundary estimation network with the generated labels and perform instance segmentation with the network. Our evaluation on a public dataset demonstrated that the proposed method achieves the best performance among the several weakly supervised methods we compared.
**********

DRAMA: Joint Risk Localization and Captioning in Driving

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, Jiachen Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1043-1052

Considering the functionality of situational awareness in safety-critical automation systems, the perception of risk in driving scenes and its explainability is of particular importance for autonomous and cooperative driving. Toward this goal, this paper proposes a new research direction of joint risk localization in driving scenes and its risk explanation as a natural language description. Due to the lack of standard benchmarks, we collected a large-scale dataset, DRAMA (Driving Risk Assessment Mechanism with A captioning module), which consists of 17,785 interactive driving scenarios collected in Tokyo, Japan. Our DRAMA dataset accommodates video- and object-level questions on driving risks with associated important objects to achieve the goal of visual captioning as a free-form language description utilizing closed and open-ended responses for multi-level questions, which can be used to evaluate a range of visual captioning capabilities in driving scenarios. We make this data available to the community for further research. Using DRAMA, we explore multiple facets of joint risk localization and captioning in interactive driving scenarios. In particular, we benchmark various multi-task prediction architectures and provide a detailed analysis of joint risk localization and risk captioning. The data set is available at https://usa.honda-ri.com/drama

**********

Vision Transformer for NeRF-Based View Synthesis From a Single Input Image

Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, Ravi Ramamoorthi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 806-815

Although neural radiance fields (NeRF) have shown impressive advances in novel view synthesis, most methods require multiple input images of the same scene with accurate camera poses. In this work, we seek to substantially reduce the inputs to a single unposed image. Existing approaches using local image features to reconstruct a 3D object often render blurry predictions at viewpoints distant from the source view. To address this, we propose to leverage both the global and local features to form an expressive 3D representation. The global features are learned from a vision transformer, while the local features are extracted from a 2D convolutional network. To synthesize a novel view, we train a multi-layer perceptron (MLP) network conditioned on the learned 3D representation to perform volume rendering. This novel 3D representation allows the network to reconstruct unseen regions without enforcing constraints like symmetry or canonical coordinate systems. Our method renders novel views from just a single input image, and generalizes across multiple object categories using a single model. Quantitative and qualitative evaluations demonstrate that the proposed method achieves state-of-the-art performance and renders richer details than existing approaches.

**********

DBCE: A Saliency Method for Medical Deep Learning Through Anatomically-Consistent Free-Form Deformations

Joshua Peters, Léo Lebrat, Rodrigo Santa Cruz, Aaron Nicolson, Gregg Belous, Salamata Konate, Parnesh Raniga, Vincent Dore, Pierrick Bourgeat, Jurgen Mejan-Fripp, Clinton Fookes, Olivier Salvado; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1959-1969

Deep learning models are powerful tools for addressing challenging medical imaging problems. However, for an ever-growing range of applications, interpreting a model's prediction remains non-trivial. Understanding decisions made by black-box algorithms is critical, and assessing their fairness and susceptibility to bias is a key step towards healthcare deployment. In this paper, we propose DBCE (Deformation Based Counterfactual Explainability). We optimise a diffeomorphic transformation that deforms a given input image to change the prediction of the model. This provides anatomically meaningful saliency maps indicating tissue atrophy and expansion, which can be easily interpreted by clinicians. In our test case, DBCE replicates the transition of a patient from healthy control (HC) to Alzheimer's disease (AD). We benchmark DBCE against three commonly used saliency meth

ods. We show that it provides more meaningful saliency maps when applied to one subject and disease-consistent atrophy patterns when used over a larger cohort. In addition, our method fulfils a recent sanity check and is repeatable for different model initialisations in contrast to classical sensitivity-based methods.
**********

EventPoint: Self-Supervised Interest Point Detection and Description for Event-Based Camera

Ze Huang, Li Sun, Cheng Zhao, Song Li, Songzhi Su; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5396-5405

This paper proposes a self-supervised learned local detector and descriptor, called EventPoint, for event stream/camera tracking and registration. Event-based cameras have grown in popularity because of their biological inspiration and low power consumption. Despite this, applying local features directly to the event stream is difficult due to its peculiar data structure. We propose a new time-surface-like event stream representation method called Tencode. The event stream data processed by Tencode can obtain the pixel-level positioning of interest points while also simultaneously extracting descriptors through a neural network. Instead of using costly and unreliable manual annotation, our network leverages the prior knowledge of local feature extraction on color images and conducts self-supervised learning via homographic and spatio-temporal adaptation. To the best of our knowledge, our proposed method is the first research on event-based local features learning using a deep neural network. We provide comprehensive experiments of feature point detection and matching, and three public datasets are used for evaluation (i.e. DSEC, N-Caltech101, and HVGA ATIS Corner Dataset). The experimental findings demonstrate that our method outperforms SOTA in terms of feature point detection and description.
**********

Leveraging Off-the-Shelf Diffusion Model for Multi-Attribute Fashion Image Manipulation

Chaerin Kong, DongHyeon Jeon, Ohjoon Kwon, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 848-857

Fashion attribute editing is a task that aims to convert the semantic attributes of a given fashion image while preserving the irrelevant regions. Previous works typically employ conditional GANs where the generator explicitly learns the target attributes and directly execute the conversion. These approaches, however, are neither scalable nor generic as they operate only with few limited attributes and a separate generator is required for each dataset or attribute set. Inspired by the recent advancement of diffusion models, we explore the classifier-guided diffusion that leverages the off-the-shelf diffusion model pretrained on general visual semantics such as Imagenet. In order to achieve a generic editing pipeline, we pose this as multi-attribute image manipulation task, where the attribute ranges from item category, fabric, pattern to collar and neckline. We empirically show that conventional methods fail in our challenging setting, and study efficient adaptation scheme that involves recently introduced attention-pooling technique to obtain a multi-attribute classifier guidance. Based on this, we present a mask-free fashion attribute editing framework that leverages the classifier logits and the cross-attention map for manipulation. We empirically demonstrate that our framework achieves convincing sample quality and attribute alignments.
**********

X-NeRF: Explicit Neural Radiance Field for Multi-Scene 360deg Insufficient RGB-D Views

Haoyi Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5766-5775

Neural Radiance Fields (NeRFs), despite their outstanding performance on novel view synthesis, often need dense input views. Many papers train one model for each scene respectively and few of them explore incorporating multi-modal data into this problem. In this paper, we focus on a rarely discussed but important setting: can we train one model that can represent multiple scenes, with 360deg insuf

ficient views and RGB-D images? We refer insufficient views to few extremely spa rse and almost non-overlapping views. To deal with it, X-NeRF, a fully explicit approach which learns a general scene completion process instead of a coordinate -based mapping, is proposed. Given a few insufficient RGB-D input views, X-NeRF first transforms them to a sparse point cloud tensor and then applies a 3D spars e generative Convolutional Neural Network (CNN) to complete it to an explicit ra diance field whose volumetric rendering can be conducted fast without running ne tworks during inference. To avoid overfitting, besides common rendering loss, we apply perceptual loss as well as view augmentation through random rotation on p oint clouds. The proposed methodology significantly out-performs previous implic it methods in our setting, indicating the great potential of proposed problem an d approach. Codes and data are available at https://github.com/HaoyiZhu/XNeRF.
**********

Attend Who Is Weak: Pruning-Assisted Medical Image Localization Under Sophistica ted and Implicit Imbalances

Ajay Jaiswal, Tianlong Chen, Justin F. Rousseau, Yifan Peng, Ying Ding, Zhangyan g Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Compute r Vision (WACV), 2023, pp. 4987-4996

Deep neural networks (DNNs) have rapidly become a de facto choice to medical ima ge understanding tasks. However, DNNs are notoriously fragile to the class imbal ance in image classification. We further point out that such imbalance fragility can be amplified when it comes to more sophisticated tasks such as pathology lo calization, as imbalances in such problems can have highly complex and often imp licit forms of presence. For example, different pathology can have different siz es or colors (w.r.t.the background), different underlying demographic distributi ons, and in general different difficulty levels to recognize, even in a meticulo usly curated balanced distribution of training data. In this paper, we propose t o use pruning to automatically and adaptively identify hard-to-learn (HTL) train ing samples, and improve pathology localization by attending them explicitly, du ring training in supervised, semi-supervised, and weakly-supervised settings. Ou r main inspiration is drawn from the recent finding that deep classification mod els have difficult-to-memorize samples and those may be effectively exposed thro ugh network pruning - and we extend such observation beyond classification for t he first time. We also present interesting demographic analysis which illustrate s HTLs ability to capture complex demographic imbalances. Our extensive experime nts on the Skin Lesion Localization task in multiple training settings by paying additional attention to HTLs show significant improvement of localization perfo rmance by 2-3%.
**********

Dynamic Neural Portraits

Michail Christos Doukas, Stylianos Ploumpis, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4073-4083

We present Dynamic Neural Portraits, a novel approach to the problem of full-hea d reenactment. Our method generates photo-realistic video portraits by explicitl y controlling head pose, facial expressions and eye gaze. Our proposed architect ure is different from existing methods that rely on GAN-based image-to-image tra nslation networks for transforming renderings of 3D faces into photo-realistic i mages. Instead, we build our system upon a 2D coordinate-based MLP with controll able dynamics. Our intuition to adopt a 2D-based representation, as opposed to r ecent 3D NeRF-like systems, stems from the fact that video portraits are capture d by monocular stationary cameras, therefore, only a single viewpoint of the sce ne is available. Primarily, we condition our generative model on expression blen dshapes, nonetheless, we show that our system can be successfully driven by audi o features as well. Our experiments demonstrate that the proposed method is 270 times faster than recent NeRF-based reenactment methods, with our networks achie ving speeds of 24 fps for resolutions up to 1024x1024, while outperforming prior works in terms of visual quality.
**********

SAT: Scale-Augmented Transformer for Person Search

Mustansar Fiaz, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4820-4829

Person search is a challenging computer vision problem where the objective is to simultaneously detect and reidentify a target person from the gallery of whole scene images captured from multiple cameras. Here, the challenges related to underlying detection and re-identification tasks need to be addressed along with a joint optimization of these two tasks. In this paper, we propose a three-stage cascaded Scale-Augmented Transformer (SAT) person search framework. In the three-stage design of our SAT framework, the first stage performs person detection whereas the last two stages performs both detection and re-identification. Considering the contradictory nature of detection and identification, in the last two stages, we introduce separate norm feature embeddings for the two tasks to reconcile the relationship between them in a joint person search model. Our SAT framework benefits from the attributes of convolutional neural networks and transformers by introducing a convolutional encoder and a scale modulator within each stage. Here, the convolutional encoder increases the generalization ability of the model whereas the scale modulator performs context aggregation at different granularity levels to aid in handling pose/scale variations within a region of interest. To further improve the performance during occlusion, we apply shifting augmentation operations at each granularity level within the scale modulator. Experimental results on challenging CUHK-SYSU [35] and PRW [47] datasets demonstrate the favorable performance of our method compared to state-of-the-art methods. Our source code and trained models are available at this https URL.

**********

Self-Supervised Learning With Masked Image Modeling for Teeth Numbering, Detection of Dental Restorations, and Instance Segmentation in Dental Panoramic Radiographs

Amani Almalki, Longin Jan Latecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5594-5603

The computer-assisted radiologic informative report is currently emerging in dental practice to facilitate dental care and reduce time consumption in manual panoramic radiographic interpretation. However, the amount of dental radiographs for training is very limited, particularly from the point of view of deep learning. This study aims to utilize recent self-supervised learning methods like SimMIM and UM-MAE to increase the model efficiency and understanding of the limited number of dental radiographs. We use the Swin Transformer for teeth numbering, detection of dental restorations, and instance segmentation tasks. To the best of our knowledge, this is the first study that applied self-supervised learning methods to Swin Transformer on dental panoramic radiographs. Our results show that the SimMIM method obtained the highest performance of 90.4% and 88.9% on detecting teeth and dental restorations and instance segmentation, respectively, increasing the average precision by 13.4 and 12.8 over the random initialization baseline. Moreover, we augment and correct the existing dataset of panoramic radiographs. The code and the dataset are available at https://github.com/AmaniHAlmalki/DentalMIM.

**********

Domain Invariant Vision Transformer Learning for Face Anti-Spoofing

Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, Chu-Song Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6098-6107

Existing face anti-spoofing (FAS) models have achieved high performance on specific datasets. However, for the application of real-world systems, the FAS model should generalize to the data from unknown domains rather than only achieve good results on a single baseline. As vision transformer models have demonstrated astonishing performance and strong capability in learning discriminative information, we investigate applying transformers to distinguish the face presentation attacks over unknown domains. In this work, we propose the Domain-invariant Vision Transformer (DiVT) for FAS, which adopts two losses to improve the generalizability of the vision transformer. First, a concentration loss is employed to learn

a domain-invariant representation that aggregates the features of real face data. Second, a separation loss is utilized to union each type of attack from different domains. The experimental results show that our proposed method achieves state-of-the-art performance on the protocols of domain-generalized FAS tasks. Compared to previous domain generalization FAS models, our proposed method is simpler but more effective.

**********

CNN2Graph: Building Graphs for Image Classification

Vivek Trivedy, Longin Jan Latecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1-11

Neural Network classifiers generally operate via the i.i.d. assumption where examples are passed through independently during training. We propose CNN2GNN and CNN2Transformer which instead leverage inter-example information for classification. We use Graph Neural Networks (GNNs) to build a latent space bipartite graph and compute cross-attention scores between input images and a proxy set. Our approach addresses several challenges of existing methods. Firstly, it is end-to-end differentiable despite the generally discrete nature of graph construction. Secondly, it allows inductive inference at no extra cost. Thirdly, it presents a simple method to construct graphs from arbitrary datasets that captures both example level and class level information. Finally, it addresses the proxy collapse problem by combining contrastive and cross-entropy losses rather than separate clustering algorithms. Our results increase classification performance over baseline experiments and outperform other methods. We also conduct an empirical investigation showing that Transformer style attention scales better than GAT attention with dataset size.

**********

Automated Line Labelling: Dataset for Contour Detection and 3D Reconstruction

Hari Santhanam, Nehal Doiphode, Jianbo Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3136-3145

Understanding the finer details of a 3D object, its contours, is the first step toward a physical understanding of an object. Many real-world application domains require adaptable 3D object shape recognition models, usually with little training data. For this purpose, we develop the first automatically generated contour labeled dataset, bypassing manual human labeling. Using this dataset, we study the performance of current state-of-the-art instance segmentation algorithms on detecting and labeling the contours. We produce promising visual results with accurate contour prediction and labeling. We demonstrate that our finely labeled contours can help downstream tasks in computer vision, such as 3D reconstruction from a 2D image.

**********

HIME: Efficient Headshot Image Super-Resolution With Multiple Exemplars

Xiaoyu Xiang, Jon Morton, Fitsum A. Reda, Lucas D. Young, Federico Perazzi, Rakesh Ranjan, Amit Kumar, Andrea Colaco, Jan P. Allebach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1694-1704

A promising direction for recovering the lost information in low-resolution headshot images is utilizing a set of high-resolution exemplars from the same identity. Complementary images in the reference set can improve the generated headshot quality across many different views and poses. However, it is challenging to make the best use of multiple exemplars: the quality and alignment of each exemplar cannot be guaranteed. Using low-quality and mismatched images as references will impair the output results. To overcome these issues, we propose the efficient Headshot Image Super-Resolution with Multiple Exemplars network (HIME) method. Compared with previous methods, our network can effectively handle the misalignment between the input and the reference without requiring facial priors and learn the aggregated reference set representation in an end-to-end manner. Furthermore, to reconstruct more detailed facial features, we propose a correlation loss that provides a rich representation of the local texture in a controllable spatial range. Experimental results demonstrate that the proposed framework not only has significantly fewer computation cost than recent exemplar-guided methods but

also achieves better qualitative and quantitative performance.
**********
Frequency-Aware Self-Supervised Monocular Depth Estimation
Xingyu Chen, Thomas H. Li, Ruonan Zhang, Ge Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5808-5817
We present two versatile methods to generally enhance self-supervised monocular depth estimation (MDE) models. The high generalizability of our methods is achieved by solving the fundamental and ubiquitous problems in photometric loss function. In particular, from the perspective of spatial frequency, we first propose Ambiguity-Masking to suppress the incorrect supervision under photometric loss at specific object boundaries, the cause of which could be traced to pixel-level ambiguity. Second, we present a novel frequency-adaptive Gaussian low-pass filter, designed to robustify the photometric loss in high-frequency regions. We are the first to propose blurring images to improve depth estimators with an interpretable analysis. Both modules are lightweight, adding no parameters and no need to manually change the network structures. Experiments show that our methods provide performance boosts to a large number of existing models, including those who claimed state-of-the-art, while introducing no extra inference computation at all.
**********
Dense Prediction With Attentive Feature Aggregation
Yung-Hsu Yang, Thomas E. Huang, Min Sun, Samuel Rota Bulò, Peter Kontschieder, Fisher Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 97-106
Aggregating information from features across different layers is essential for dense prediction models. Despite its limited expressiveness, vanilla feature concatenation dominates the choice of aggregation operations. In this paper, we introduce Attentive Feature Aggregation (AFA) to fuse different network layers with more expressive non-linear operations. AFA exploits both spatial and channel attention to compute weighted averages of the layer activations. Inspired by neural volume rendering, we further extend AFA with Scale-Space Rendering (SSR) to perform a late fusion of multi-scale predictions. AFA is applicable to a wide range of existing network designs. Our experiments show consistent and significant improvements on challenging semantic segmentation benchmarks, including Cityscapes and BDD100K at negligible computational and parameter overhead. In particular, AFA improves the performance of the Deep Layer Aggregation (DLA) model by nearly 6% mIoU on Cityscapes. Our experimental analyses show that AFA learns to progressively refine segmentation maps and improve boundary details, leading to new state-of-the-art results on boundary detection benchmarks on NYUDv2 and BSDS500.
**********
Interacting Hand-Object Pose Estimation via Dense Mutual Attention
Rong Wang, Wei Mao, Hongdong Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5735-5745
3D hand-object pose estimation is the key to the success of many computer vision applications. The main focus of this task is to effectively model the interaction between the hand and an object. To this end, existing works either rely on interaction constraints in a computationally-expensive iterative optimization, or consider only a sparse correlation between sampled hand and object keypoints. In contrast, we propose a novel dense mutual attention mechanism that is able to model fine-grained dependencies between the hand and the object. Specifically, we first construct the hand and object graphs according to their mesh structures. For each hand node, we aggregate features from every object node by the learned attention and vice versa for each object node. Thanks to such dense mutual attention, our method is able to produce physically plausible poses with high quality and real-time inference speed. Extensive quantitative and qualitative experiments on large benchmark datasets show that our method outperforms state-of-the-art methods. The code is available at https://github.com/rongakowang/DenseMutualAttention.git.
**********
Detection Recovery in Online Multi-Object Tracking With Sparse Graph Tracker

Jeongseok Hyun, Myunggu Kang, Dongyoon Wee, Dit-Yan Yeung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4850-4859

In existing joint detection and tracking methods, pairwise relational features are used to match previous tracklets to current detections. However, the features may not be discriminative enough for a tracker to identify a target from a large number of detections. Selecting only high-scored detections for tracking may lead to missed detections whose confidence score is low. Consequently, in the online setting, this results in disconnections of tracklets which cannot be recovered. In this regard, we present Sparse Graph Tracker (SGT), a novel online graph tracker using higher-order relational features which are more discriminative by aggregating the features of neighboring detections and their relations. SGT converts video data into a graph where detections, their connections, and the relational features of two connected nodes are represented by nodes, edges, and edge features, respectively. The strong edge features allow SGT to track targets with tracking candidates selected by top-K scored detections with large K. As a result, even low-scored detections can be tracked, and the missed detections are also recovered. The robustness of K value is shown through the extensive experiments. In the MOT16/17/20 and HiEve Challenge, SGT outperforms the state-of-the-art trackers with real-time inference speed. Especially, a large improvement in MOTA is shown in the MOT20 and HiEve Challenge. Code is available at https://github.com/HYUNJS/SGT.

**********

Analysis of Master Vein Attacks on Finger Vein Recognition Systems

Huy H. Nguyen, Trung-Nghia Le, Junichi Yamagishi, Isao Echizen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1900-1908

Finger vein recognition (FVR) systems have been commercially used, especially in ATMs, for customer verification. Thus, it is essential to measure their robustness against various attack methods, especially when a hand-crafted FVR system is used without any countermeasure methods. In this paper, we are the first in the literature to introduce master vein attacks in which we craft a vein-looking image so that it can falsely match with as many identities as possible by the FVR systems. We present two methods for generating master veins for use in attacking these systems. The first uses an adaptation of the latent variable evolution algorithm with a proposed generative model (a multi-stage combination of beta-VAE and WGAN-GP models). The second uses an adversarial machine learning attack method to attack a strong surrogate CNN-based recognition system. The two methods can be easily combined to boost their attack ability. Experimental results demonstrated that the proposed methods alone and together achieved false acceptance rates up to 73.29% and 88.79%, respectively, against Miura's hand-crafted FVR system. We also point out that Miura's system is easily compromised by non-vein-looking samples generated by a WGAN-GP model with false acceptance rates up to 94.21%. The results raise the alarm about the robustness of such systems and

**********

Image Completion With Heterogeneously Filtered Spectral Hints

Xingqian Xu, Shant Navasardyan, Vahram Tadevosyan, Andranik Sargsyan, Yadong Mu, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4591-4601

Image completion with large-scale free-form missing regions is one of the most challenging tasks for the computer vision community. While researchers pursue better solutions, drawbacks such as pattern unawareness, blurry textures, and structure distortion remain noticeable, and thus leave space for improvement. To overcome these challenges, we propose a new StyleGAN-based image completion network, Spectral Hint GAN (SH-GAN), inside which a carefully designed spectral processing module, Spectral Hint Unit, is introduced. We also propose two novel 2D spectral processing strategies, Heterogeneous Filtering, and Gaussian Split that well-fit modern deep learning models and may further be extended to other tasks. From our inclusive experiments, we demonstrate that our model can reach FID scores of 3.4134 and 7.0277 on the benchmark datasets FFHQ and Places2, and therefore o

utperforms prior works and reaches a new state-of-the-art. We also prove the eff
ectiveness of our design via ablation studies, from which one may notice that th
e aforementioned challenges, i.e. pattern unawareness, blurry textures, and stru
cture distortion, can be noticeably resolved. Our code will be open-sourced at:
https://github.com/SHI-Labs/SH-GAN.
**********

Split To Learn: Gradient Split for Multi-Task Human Image Analysis
Weijian Deng, Yumin Suh, Xiang Yu, Masoud Faraki, Liang Zheng, Manmohan Chandrak
er; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi
sion (WACV), 2023, pp. 4351-4360
This paper presents an approach to train a unified deep network that simultaneou
sly solves multiple human-related tasks. A multi-task framework is favorable for
 sharing information across tasks under restricted computational resources. Howe
ver, tasks not only share information but may also compete for resources and con
flict with each other, making the optimization of shared parameters difficult an
d leading to suboptimal performance. We propose a simple but effective training
scheme called GradSplit that alleviates this issue by utilizing asymmetric inter
-task relations. Specifically, at each convolution module, it splits features in
to T groups for T tasks and trains each group only using the gradient back-propa
gated from the task losses with which it does not have conflicts. During trainin
g, we apply GradSplit to a series of convolution modules. As a result, each modu
le is trained to generate a set of task-specific features using the shared featu
res from the previous module. This enables a network to use complementary inform
ation across tasks while circumventing gradient conflicts. Experimental results
show that GradSplit achieves a better accuracy-efficiency trade-off than existin
g methods. It minimizes accuracy drop caused by task conflicts while significant
ly saving compute resources in terms of both FLOPs and memory at inference. We f
urther show that GradSplit achieves higher cross-dataset accuracy compared to si
ngle-task and other multi-task networks.
**********

TransPillars: Coarse-To-Fine Aggregation for Multi-Frame 3D Object Detection
Zhipeng Luo, Gongjie Zhang, Changqing Zhou, Tianrui Liu, Shijian Lu, Liang Pan;
Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision
 (WACV), 2023, pp. 4230-4239
3D object detection using point clouds has attracted increasing attention due to
 its wide applications in autonomous driving and robotics. However, most existin
g studies focus on single point cloud frames without harnessing the temporal inf
ormation in point cloud sequences. In this paper, we design TransPillars, a nove
l transformer-based feature aggregation technique that exploits temporal feature
s of consecutive point cloud frames for multi-frame 3D object detection. TransPi
llars aggregates spatial-temporal point cloud features from two perspectives. Fi
rst, it fuses voxel-level features directly from multi-frame feature maps instea
d of pooled instance features to preserve instance details with contextual infor
mation that are essential to accurate object localization. Second, it introduces
 a hierarchical coarse-to-fine strategy to fuse multi-scale features progressive
ly to effectively capture the motion of moving objects and guide the aggregation
 of fine features. Besides, a variant of deformable transformer is introduced to
 improve the effectiveness of cross-frame feature matching. Extensive experiment
s show that our proposed TransPillars achieves state-of-art performance as compa
red to existing multi-frame detection approaches.
**********

Body Part-Based Representation Learning for Occluded Person Re-Identification
Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi; Proceedings of the
 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp.
 1613-1623
Occluded person re-identification (ReID) is a person retrieval task which aims a
t matching occluded person images with holistic ones. For addressing occluded Re
ID, part-based methods have been shown beneficial as they offer fine-grained inf
ormation and are well suited to represent partially visible human bodies. Howeve
r, training a part-based model is a challenging task for two reasons. Firstly, i

ndividual body part appearance is not as discriminative as global appearance (tw
o distinct IDs might have the same local appearance), this means standard ReID t
raining objectives using identity labels are not adapted to local feature learni
ng. Secondly, ReID datasets are not provided with human topographical annotation
s. In this work, we propose BPBreID, a body part-based ReID model for solving th
e above issues. We first design two modules for predicting body part attention m
aps and producing body part-based features of the ReID target. We then propose G
iLt, a novel training scheme for learning part-based representations that is rob
ust to occlusions and non-discriminative local appearance. Extensive experiments
 on popular holistic and occluded datasets show the effectiveness of our propose
d method, which outperforms state-of-the-art methods by 0.7% mAP and 5.6% rank-1
 accuracy on the challenging Occluded-Duke dataset. Our code is available at htt
ps://github.com/VlSomers/bpbreid.
**********

Generative Range Imaging for Learning Scene Priors of 3D LiDAR Data
Kazuto Nakashima, Yumi Iwashita, Ryo Kurazume; Proceedings of the IEEE/CVF Winte
r Conference on Applications of Computer Vision (WACV), 2023, pp. 1256-1266
3D LiDAR sensors are indispensable for the robust vision of autonomous mobile ro
bots. However, deploying LiDAR-based perception algorithms often fails due to a
domain gap from the training environment, such as inconsistent angular resolutio
n and missing properties. Existing studies have tackled the issue by learning in
ter-domain mapping, while the transferability is constrained by the training con
figuration and the training is susceptible to peculiar lossy noises called ray-d
rop. To address the issue, this paper proposes a generative model of LiDAR range
 images applicable to the data-level domain transfer. Motivated by the fact that
 LiDAR measurement is based on point-by-point range imaging, we train an implici
t image representation-based generative adversarial networks along with a differ
entiable ray-drop effect. We demonstrate the fidelity and diversity of our model
 in comparison with the point-based and image-based state-of-the-art generative
models. We also showcase upsampling and restoration applications. Furthermore, w
e introduce a Sim2Real application for LiDAR semantic segmentation. We demonstra
te that our method is effective as a realistic ray-drop simulator and outperform
s state-of-the-art methods.
**********

Placing Human Animations Into 3D Scenes by Learning Interaction- and Geometry-Dr
iven Keyframes
James F. Mullen, Divya Kothandaraman, Aniket Bera, Dinesh Manocha; Proceedings o
f the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023
, pp. 300-310
We present a novel method for placing a 3D human animation into a 3D scene while
 maintaining any human-scene interactions in the animation. We use the notion of
 computing the most important meshes in the animation for the interaction with t
he scene, which we call "keyframes." These keyframes allow us to better optimize
 the placement of the animation into the scene such that interactions in the ani
mations (standing, laying, sitting, etc.) match the affordances of the scene (e.
g., standing on the floor or laying in a bed). We compare our method, which we c
all PAAK, with prior approaches, including POSA, PROX ground truth, and a motion
 synthesis method, and highlight the benefits of our method with a perceptual st
udy. Human raters preferred our PAAK method over the PROX ground truth data 64.6
% of the time. Additionally, in direct comparisons, the raters preferred PAAK ov
er competing methods including 61.5% compared to POSA. Our project website is av
ailable at https://gamma.umd.edu/paak/.
**********

Rebalancing Gradient To Improve Self-Supervised Co-Training of Depth, Odometry a
nd Optical Flow Predictions
Marwane Hariat, Antoine Manzanera, David Filliat; Proceedings of the IEEE/CVF Wi
nter Conference on Applications of Computer Vision (WACV), 2023, pp. 1267-1276
We present CoopNet, an approach that improves the cooperation of co-trained netw
orks by dynamically adapting the apportionment of gradient, to ensure equitable
learning progress. It is applied to motion-aware self-supervised prediction of d

epth maps, by introducing a new hybrid loss, based on a distribution model of ph
oto-metric reconstruction errors made by, on the one hand the depth + odometry p
aired networks, and on the other hand the optical flow network. This model essen
tially assumes that the pixels from moving objects (that must be discarded for t
raining depth and odometry), correspond to those where the two reconstructions s
trongly disagree. We justify this model by theoretical considerations and experi
mental evidences, and show that its implementation improves or is comparable to
the state of the art in depth, odometry and optical flow predictions. Our code i
s available here: https://github.com/mhariat/CoopNet.
**********

RADIANT: Better rPPG Estimation Using Signal Embeddings and Transformer
Anup Kumar Gupta, Rupesh Kumar, Lokendra Birla, Puneet Gupta; Proceedings of the
 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp.
 4976-4986
Remote photoplethysmography can provide non-contact heart rate (HR) estimation b
y analyzing the skin color variations obtained from face videos. These variation
s are subtle, imperceptible to human eyes, and easily affected by noise. Existin
g deep learning-based rPPG estimators are incompetent due to three reasons. Firs
tly, they suppress the noise by utilizing information from the whole face even t
hough different facial regions contain different noise characteristics. Secondly
, local noise characteristics inherently affect the convolutional neural network
 (CNN) architectures. Lastly, the CNN sequential architectures fail to preserve
long temporal dependencies. To address these issues, we propose RADIANT, that is
, rPPG estimation using Signal Embeddings and Transformer. Our Transformer utili
zes a multi-head attention mechanism that facilitates the feature subspace learn
ing to extract the multiple correlations among the color variations correspondin
g to the periodic pulse. Also, its global information processing ability helps t
o suppress local noise characteristics. Apart from Transformer, we propose novel
 signal embedding to enhance the rPPG feature representation and suppress noise.
 We have also improved the generalization of our architecture by adding a new tr
aining set. To this end, the effectiveness of synthetic temporal signals and dat
a augmentations were explored. Experiments on extensively utilized UBFC-rPPG and
 COHFACE datasets demonstrate that our architecture outperforms previous well-kn
own architectures. The implementation will be made publicly available upon paper
 acceptance.
**********

Marker-Removal Networks To Collect Precise 3D Hand Data for RGB-Based Estimation
 and Its Application in Piano
Erwin Wu, Hayato Nishioka, Shinichi Furuya, Hideki Koike; Proceedings of the IEE
E/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 297
7-2986
Hand pose analysis is a key step to understanding dexterous hand performances of
 many high-level skills, such as playing the piano. Currently, most accurate han
d tracking systems are using fabric-/marker-based sensing that potentially distu
rbs users' performance. On the other hand, markerless computer vision-based meth
ods rely on a precise bare-hand dataset for training, which is difficult to obta
in. In this paper, we collect a large-scale high precision 3D hand pose dataset
with a small workload using a novel marker-removal network (MR-Net). The propose
d MR-Net translates the marked-hand images to realistic bare-hand images, and th
e corresponding 3D postures are captured by a motion capture system thus few man
ual annotations are required. A baseline estimation network PiaNet is introduced
 and we report the accuracy of various metrics together with a blind qualitative
 test to show the practical effect.
**********

Cross-Identity Video Motion Retargeting With Joint Transformation and Synthesis
Haomiao Ni, Yihao Liu, Sharon X. Huang, Yuan Xue; Proceedings of the IEEE/CVF Wi
nter Conference on Applications of Computer Vision (WACV), 2023, pp. 412-422
In this paper, we propose a novel dual-branch Transformation-Synthesis network (
TS-Net), for video motion retargeting. Given one subject video and one driving v
ideo, TS-Net can produce a new plausible video with the subject appearance of th

e subject video and motion pattern of the driving video. TS-Net consists of a wa rp-based transformation branch and a warp-free synthesis branch. The novel desig n of dual branches combines the strengths of deformation-grid-based transformati on and warp-free generation for better identity preservation and robustness to o cclusion in the synthesized videos. A mask-aware similarity module is further in troduced to the transformation branch to reduce computational overhead. Experime ntal results on face and dance datasets show that TS-Net achieves better perform ance in video motion retargeting than several state-of-the-art models as well as its single-branch variants. Our code is available at https://github.com/nihaomi ao/WACV23_TSNet.

**********

Learning Across Domains and Devices: Style-Driven Source-Free Domain Adaptation in Clustered Federated Learning

Donald Shenaj, Eros Fanì, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro Zanuttigh, Barbara Caputo; Proceedings of the I EEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4 44-454

Federated Learning (FL) has recently emerged as a possible way to tackle the dom ain shift in real-world Semantic Segmentation (SS) without compromising the priv ate nature of the collected data. However, most of the existing works on FL unre alistically assume labeled data in the remote clients. Here we propose a novel t ask (FFREEDA) in which the clients' data is unlabeled and the server accesses a source labeled dataset for pre-training only. To solve FFREEDA, we propose LADD, which leverages the knowledge of the pre-trained model by employing self-superv ision with ad-hoc regularization techniques for local training and introducing a novel federated clustered aggregation scheme based on the clients' style. Our e xperiments show that our algorithm is able to efficiently tackle the new task ou tperforming existing approaches. The code is available at https://github.com/Ero sinho13/LADD.

**********

Learning Attention Propagation for Compositional Zero-Shot Learning

Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, D idier Stricker, Muhammad Zeshan Afzal; Proceedings of the IEEE/CVF Winter Confer ence on Applications of Computer Vision (WACV), 2023, pp. 3828-3837

Compositional zero-shot learning aims to recognize unseen compositions of seen v isual primitives of object classes and their states. While all primitives (state s and objects) are observable during training in some combination, their complex interaction makes this task especially hard. For example, wet changes the visua l appearance of a dog very differently from a bicycle. Furthermore, we argue tha t relationships between compositions go beyond shared states or objects. A clutt ered office can contain a busy table; even though these compositions don't share a state or object, the presence of a busy table can guide the presence of a clu ttered office. We propose a novel method called Compositional Attention Propagat ed Embedding (CAPE) as a solution. The key intuition to our method is that a ric h dependency structure exists between compositions arising from complex interact ions of primitives in addition to other dependencies between compositions. CAPE learns to identify this structure and propagates knowledge between them to learn class embedding for all seen and unseen compositions. In the challenging genera lized compositional zero-shot setting, we show that our method outperforms previ ous baselines to set a new state-of-the-art on three publicly available benchmar ks.

**********

Language-Free Training for Zero-Shot Video Grounding

Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, Kwanghoon Sohn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20 23, pp. 2539-2548

Given an untrimmed video and a language query depicting a specific temporal mome nt in the video, video grounding aims to localize the time interval by understan ding the text and video simultaneously. One of the most challenging issues is an extremely time- and cost-consuming annotation collection, including video capti

ons in a natural language form and their corresponding temporal regions. In this paper, we present a simple yet novel training framework for video grounding in the zero-shot setting, which learns a network with only video data without any annotation. Inspired by the recent language-free paradigm, i.e. training without language data, we train the network without compelling the generation of fake (pseudo) text queries into a natural language form. Specifically, we propose a method for learning a video grounding model by selecting a temporal interval as a hypothetical correct answer and considering the visual feature selected by our method in the interval as a language feature, with the help of the well-aligned visual-language space of CLIP. Extensive experiments demonstrate the prominence of our language-free training framework, outperforming the existing zero-shot video grounding method and even several weakly-supervised approaches with large margins on two standard datasets.

**********

Weakly-Supervised Point Cloud Instance Segmentation With Geometric Priors

Heming Du, Xin Yu, Farookh Hussain, Mohammad Ali Armin, Lars Petersson, Weihao Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4271-4280

This paper investigates how to leverage more readily acquired annotations, i.e., 3D bounding boxes instead of dense point-wise labels, for instance segmentation. We propose a Weakly-supervised point cloud Instance Segmentation framework with Geometric Priors (WISGP) that allows segmentation models to be trained with 3D bounding boxes of instances. Considering intersections among bounding boxes in a scene would result in ambiguous labels, we first group points into two sets, i.e., univocal and equivocal sets, indicating the certainty of a 3D point belonging to an instance, respectively. Specifically, 3D points with clear labels belong to the univocal set while the rest are grouped into the equivocal set. To assign reliable labels to points in the equivocal set, we design a Geometry-guided Label Propagation (GLP) scheme that progressively propagates labels to linked points based on geometric structure, e.g., polygon meshes and superpoints. Afterwards, we train an instance segmentation model with the univocal points and equivocal points labeled by GLP, and then employ it to assign pseudo labels for the remainder of the unlabeled points. Lastly, we retrain the model with all the labeled points to achieve better instance segmentation performance. Experiments on large-scale datasets ScanNet-v2 and S3DIS demonstrate that WISGP is superior to competing weakly-supervised algorithms and even on par with a few fully-supervised ones.

**********

Federated Domain Generalization for Image Recognition via Cross-Client Style Transfer

Junming Chen, Meirui Jiang, Qi Dou, Qifeng Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 361-370

Domain generalization (DG) has been a hot topic in image recognition, with a goal to train a general model that can perform well on unseen domains. Recently, federated learning (FL), an emerging machine learning paradigm to train a global model from multiple decentralized clients without compromising data privacy, brings new challenges, also new possibilities, to DG. In the FL scenario, many existing state-of-the-art (SOTA) DG methods become ineffective, because they require the centralization of data from different domains during training. In this paper, we propose a novel domain generalization method for image recognition under federated learning through cross-client style transfer (CCST) without exchanging data samples. Our CCST method can lead to more uniform distributions of source clients, and thus make each local model learn to fit the image styles of all the clients to avoid the different model biases. Two types of style (single image style and overall domain style) with corresponding mechanisms are proposed to be chosen according to different scenarios. Our style representation is exceptionally lightweight and can hardly be used for the reconstruction of the dataset. The level of diversity is also flexible to be controlled with a hyper-parameter. Our method outperforms recent SOTA DG methods on two DG benchmarks (PACS, OfficeHome) and a large-scale medical image dataset (Camelyon17) in the FL setting. Last b

ut not least, our method is orthogonal to many classic DG methods, achieving additive performance by combined utilization.
**********

BEVSegFormer: Bird's Eye View Semantic Segmentation From Arbitrary Camera Rigs
Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, Erkang Cheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5935-5943
Semantic segmentation in bird's eye view (BEV) is an important task for autonomous driving. Though this task has attracted a large amount of research efforts, it is still challenging to flexibly cope with arbitrary (single or multiple) camera sensors equipped on the autonomous vehicle. In this paper, we present BEVSegFormer, an effective transformer-based method for BEV semantic segmentation from arbitrary camera rigs. Specifically, our method first encodes image features from arbitrary cameras with a shared backbone. These image features are then enhanced by a deformable transformer-based encoder. Moreover, we introduce a BEV transformer decoder module to parse BEV semantic segmentation results. An efficient multi-camera deformable attention unit is designed to carry out the BEV-to-image view transformation. Finally, the queries are reshaped according the layout of grids in the BEV, and upsampled to produce the semantic segmentation result in a supervised manner. We evaluate the proposed algorithm on the public nuScenes dataset and a self-collected dataset. Experimental results show that our method achieves promising performance on BEV semantic segmentation from arbitrary camera rigs. We also demonstrate the effectiveness of each component via ablation study.
**********

Self-Supervised 2D/3D Registration for X-Ray to CT Image Fusion
Srikrishna Jaganathan, Maximilian Kukla, Jian Wang, Karthik Shetty, Andreas Maier; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2788-2798
Deep Learning-based 2D/3D registration enables fast, robust, and accurate X-ray to CT image fusion when large annotated paired datasets are available for training. However, the need for paired CT volume and X-ray images with ground truth registration limits the applicability in interventional scenarios. An alternative is to use simulated X-ray projections from CT volumes, thus removing the need for paired annotated datasets. Deep Neural Networks trained exclusively on simulated X-ray projections can perform significantly worse on real X-ray images due to the domain gap. We propose a self-supervised 2D/3D registration framework combining simulated training with unsupervised feature and pixel space domain adaptation to overcome the domain gap and eliminate the need for paired annotated datasets. Our framework achieves a registration accuracy of 1.83 +-1.16 mm with a high success ratio of 90.1% on real X-ray images showing a 23.9% increase in success ratio compared to reference annotation-free algorithms.
**********

Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping From Egocentric Images to Allocentric Semantics With Vision Transformers
Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4013-4022
Humans have an innate ability to sense their surroundings, as they can extract the spatial representation from the egocentric perception and form an allocentric semantic map via spatial transformation and memory updating. However, endowing mobile agents with such a spatial sensing ability is still a challenge, due to two difficulties: (1) the previous convolutional models are limited by the local receptive field, thus, struggling to capture holistic long-range dependencies during observation; (2) the excessive computational budgets required for success, often lead to a separation of the mapping pipeline into stages, resulting the entire mapping process inefficient. To address these issues, we propose an end-to-end one-stage Transformer-based framework for Mapping, termed Trans4Map. Our egocentric-to-allocentric mapping process includes three steps: (1) the efficient transformer extracts the contextual features from a batch of egocentric images; (2) the proposed Bidirectional Allocentric Memory (BAM) module projects egocentri

c features into the allocentric memory; (3) the map decoder parses the accumulated memory and predicts the top-down semantic segmentation map. In contrast, Trans4Map achieves state-of-the-art results, reducing 67.2% parameters, yet gaining a +3.25% mIoU and a +4.09% mBF1 improvements on the Matterport3D dataset.
**********

Dense Voxel Fusion for 3D Object Detection
Anas Mahmoud, Jordan S. K. Hu, Steven L. Waslander; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 663-672
Camera and LiDAR sensor modalities provide complementary appearance and geometric information useful for detecting 3D objects for autonomous vehicle applications. However, current end-to-end fusion methods are challenging to train and underperform state-of-the-art LiDAR-only detectors. Sequential fusion methods suffer from a limited number of pixel and point correspondences due to point cloud sparsity, or their performance is strictly capped by the detections of one of the modalities. Our proposed solution, Dense Voxel Fusion (DVF) is a sequential fusion method that generates multi-scale dense voxel feature representations, improving expressiveness in low point density regions. To enhance multi-modal learning, we train directly with projected ground truth 3D bounding box labels, avoiding noisy, detector-specific 2D predictions. Both DVF and the multi-modal training approach can be applied to any voxel-based LiDAR backbone. DVF ranks 3rd among published fusion methods on KITTI's 3D car detection benchmark without introducing additional trainable parameters, nor requiring stereo images or dense depth labels. In addition, DVF significantly improves 3D vehicle detection performance of voxel-based methods on the Waymo Open Dataset.
**********

Effective Invertible Arbitrary Image Rescaling
Zhihong Pan, Baopu Li, Dongliang He, Wenhao Wu, Errui Ding; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5416-5425
Great successes have been achieved using deep learning techniques for image super-resolution (SR) with fixed scales. To increase its real world applicability, numerous models have also been proposed to restore SR images with arbitrary scale factors, including asymmetric ones where images are resized to different scales along horizontal and vertical directions. Though most models are only optimized for the unidirectional upscaling task while assuming a predefined downscaling kernel for low-resolution (LR) inputs, recent models based on Invertible Neural Networks (INN) are able to increase upscaling accuracy significantly by optimizing the downscaling and upscaling cycle jointly. However, limited by the INN architecture, it is constrained to fixed integer scale factors and requires one model for each scale. Without increasing model complexity, a simple and effective invertible arbitrary rescaling network (IARN) is proposed to achieve arbitrary image rescaling by training only one model in this work. Using innovative components like position-aware scale encoding and preemptive channel splitting, the network is optimized to convert the non-invertible rescaling cycle to an effectively invertible process. It is shown to achieve a state-of-the-art (SOTA) performance in bidirectional arbitrary rescaling without compromising perceptual quality in LR outputs. It is also demonstrated to perform well on tests with asymmetric scales using the same network architecture.
**********

FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping
Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, Cristofer Englund; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3454-3463
In this work, we present a new single-stage method for subject agnostic face swapping and identity transfer, named FaceDancer. We have two major contributions: Adaptive Feature Fusion Attention (AFFA) and Interpreted Feature Similarity Regularization (IFSR). The AFFA module is embedded in the decoder and adaptively learns to fuse attribute features and features conditioned on identity information without requiring any additional facial segmentation process. In IFSR, we leverage the intermediate features in an identity encoder to preserve important attrib

utes such as head pose, facial expression, lighting, and occlusion in the target face, while still transferring the identity of the source face with high fidelity. We conduct extensive quantitative and qualitative experiments on various datasets and show that the proposed FaceDancer outperforms other state-of-the-art networks in terms of identity transfer, while having significantly better pose preservation than most of the previous methods.
**********

Few-Shot Medical Image Segmentation With Cycle-Resemblance Attention
Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, Yan Yan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2488-2497
Recently, due to the increasing requirements of medical imaging applications and the professional requirements of annotating medical images, few-shot learning has gained increasing attention in the medical image semantic segmentation field. To perform segmentation with limited number of labeled medical images, most existing studies use Prototypical Networks (PN) and have obtained compelling success. However, these approaches overlook the query image features extracted from the proposed representation network, failing to preserving the spatial connection between query and support images. In this paper, we propose a novel self-supervised few-shot medical image segmentation network and introduce a novel Cycle-Resemblance Attention (CRA) module to fully leverage the pixel-wise relation between query and support medical images. Notably, we first line up multiple attention blocks to refine more abundant relation information. Then, we present CRAPNet by integrating the CRA module with a classic prototype network, where pixel-wise relations between query and support features are well recaptured for segmentation. Extensive experiments on two different medical image datasets, e.g., abdomen MRI and abdomen CT, demonstrate the superiority of our model over existing state-of-the-art methods.
**********

Dataset Condensation With Distribution Matching
Bo Zhao, Hakan Bilen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6514-6523
Computational cost of training state-of-the-art deep models in many learning problems is rapidly increasing due to more sophisticated models and larger datasets. A recent promising direction for reducing training cost is dataset condensation that aims to replace the original large training set with a significantly smaller learned synthetic set while preserving the original information. While training deep models on the small set of condensed images can be extremely fast, their synthesis remains computationally expensive due to the complex bi-level optimization and second-order derivative computation. In this work, we propose a simple yet effective method that synthesizes condensed images by matching feature distributions of the synthetic and original training images in many sampled embedding spaces. Our method significantly reduces the synthesis cost while achieving comparable or better performance. Thanks to its efficiency, we apply our method to more realistic and larger datasets with sophisticated neural architectures and obtain a significant performance boost. We also show promising practical benefits of our method in continual learning and neural architecture search.
**********

Can Shadows Reveal Biometric Information?
Safa C. Medin, Amir Weiss, Frédo Durand, William T. Freeman, Gregory W. Wornell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 869-879
We study the problem of extracting biometric information of individuals by looking at shadows of objects cast on diffuse surfaces. We show that the biometric information leakage from shadows can be sufficient for reliable identity inference under representative scenarios via a maximum likelihood analysis. We then develop a learning-based method that demonstrates this phenomenon in real settings, exploiting the subtle cues in the shadows that are the source of the leakage without requiring any labeled real data. In particular, our approach relies on building synthetic scenes composed of 3D face models obtained from a single photograp

h of each identity. We transfer what we learn from the synthetic data to the rea l data using domain adaptation in a completely unsupervised way. Our model is ab le to generalize well to the real domain and is robust to several variations in the scenes. We report high classification accuracies in an identity classificati on task that takes place in a scene with unknown geometry and occluding objects.
**********

Neural Weight Search for Scalable Task Incremental Learning
Jian Jiang, Oya Celiktutan; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 1390-1399
Task incremental learning aims to enable a system to maintain its performance on previously learned tasks while learning new tasks, solving the problem of catas trophic forgetting. One promising approach is to build an individual network or sub-network for future tasks. However, this leads to an ever-growing memory due to saving extra weights for new tasks and how to address this issue has remained an open problem in task incremental learning. In this paper, we introduce a nov el Neural Weight Search technique that designs a fixed search space where the op timal combinations of frozen weights can be searched to build new models for nov el tasks in an end-to-end manner, resulting in a scalable and controllable memor y growth. Extensive experiments on two benchmarks, i.e., Split-CIFAR-100 and CUB -to-Sketches, show our method achieves state-of-the-art performance with respect to both average inference accuracy and total memory cost.
**********

ReEnFP: Detail-Preserving Face Reconstruction by Encoding Facial Priors
Yasheng Sun, Jiangke Lin, Hang Zhou, Zhiliang Xu, Dongliang He, Hideki Koike; Pr oceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision ( WACV), 2023, pp. 6118-6128
We address the problem of face modeling, which is still challenging in achieving high-quality reconstruction results efficiently. Neither previous regression-ba sed nor optimization-based frameworks could well balance between the facial reco nstruction fidelity and efficiency. We notice that the large amount of in-the-wi ld facial images contain diverse appearance information, however, their underlyi ng knowledge is not fully exploited for face modeling. To this end, we propose o ur Reconstruction by Encoding Facial Priors (ReEnFP) pipeline to exploit the pot ential of unconstrained facial images for further improvement. Our key is to enc ode generative priors learned by a style-based texture generator on unconstraine d data for fast and detail-preserving face reconstruction. With our texture gene rator pre-trained using a differentiable renderer, faces could be encoded to its latent space as opposed to the time-consuming optimization-based inversion. Our generative prior encoding is further enhanced with a pyramid fusion block for a daptive integration of input spatial information. Extensive experiments show tha t our method reconstructs photo-realistic facial textures and geometric details with precise identity recovery.
**********

Certified Defense for Content Based Image Retrieval
Kazuya Kakizaki, Kazuto Fukuchi, Jun Sakuma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4561-4570
This paper develops a certified defense for deep neural network (DNN) based cont ent based image retrieval (CBIR) against adversarial examples (AXs). Previous wo rks put their effort into certified defense for classification to improve certif ied robustness, which guarantees that no AX to cause misclassification exists ar ound the sample. Such certified defense, however, could not be applied to CBIR d irectly because the goals of adversarial attack against classification and CBIR are completely different. To develop the certified defense for CBIR, we first de fine new certified robustness of CBIR, which guarantees that no AX that changes the ranking of CBIR exists around the query or candidate images. Then, we propos e computationally tractable verification algorithms that verify whether the cert ified robustness of CBIR is achieved by utilizing upper and lower bounds of dist ances between feature representations of perturbed and non-perturbed images. Fin ally, we propose new objective functions for training feature extraction DNNs th at increases the number of inputs that satisfy the certified robustness of CBIR

by tightening the upper and lower bounds. Experimental results show that our obj
ective functions significantly improve the certified robustness of CBIR than exi
sting methods.
**********

Refign: Align and Refine for Adaptation of Semantic Segmentation to Adverse Cond
itions
David Brüggemann, Christos Sakaridis, Prune Truong, Luc Van Gool; Proceedings of
 the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023,
 pp. 3174-3184
Due to the scarcity of dense pixel-level semantic annotations for images recorde
d in adverse visual conditions, there has been a keen interest in unsupervised d
omain adaptation (UDA) for the semantic segmentation of such images. UDA adapts
models trained on normal conditions to the target adverse-condition domains. Mea
nwhile, multiple datasets with driving scenes provide corresponding images of th
e same scenes across multiple conditions, which can serve as a form of weak supe
rvision for domain adaptation. We propose Refign, a generic extension to self-tr
aining-based UDA methods which leverages these cross-domain correspondences. Ref
ign consists of two steps: (1) aligning the normal-condition image to the corres
ponding adverse-condition image using an uncertainty-aware dense matching networ
k, and (2) refining the adverse prediction with the normal prediction using an a
daptive label correction mechanism. We design custom modules to streamline both
steps and set the new state of the art for domain-adaptive semantic segmentation
 on several adverse-condition benchmarks, including ACDC and Dark Zurich. The ap
proach introduces no extra training parameters, minimal computational overhead--
during training only--and can be used as a drop-in extension to improve any give
n self-training-based UDA method. Code is available at https://github.com/brdav/
refign.
**********

ETR: An Efficient Transformer for Re-Ranking in Visual Place Recognition
Hao Zhang, Xin Chen, Heming Jing, Yingbin Zheng, Yuan Wu, Cheng Jin; Proceedings
 of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20
23, pp. 5665-5674
Visual place recognition is to estimate the geographical location of a given ima
ge, which is usually addressed by recognizing its similar reference images from
a database. The reference images are usually retrieved via similarity search usi
ng global descriptor, and the local descriptors are used to re-rank the initial
retrieved candidates. The local descriptors re-ranking can significantly improve
 the accuracy of global retrieval but comes at a high computational cost. To ach
ieve a good trade-off between accuracy and efficiency, we propose an Efficient T
ransformer for Re-ranking (ETR), utilizing both global and local descriptors to
re-rank the top candidates in a single shot. In contrast to traditional re-ranki
ng methods, we leverage self-attention to capture relationships between local de
scriptors in a single image and cross-attention to explore the similarity of the
 image pairs. We show that the proposed model can be regarded as a general re-ra
nking algorithm for significantly boosting the performance of other global-only
retrieval methods. Extensive experimental results show that our method outperfor
ms state-of-the-arts and is orders of magnitude faster in terms of computational
 efficiency.
**********

MT-DETR: Robust End-to-End Multimodal Detection With Confidence Fusion
Shih-Yun Chu, Ming-Sui Lee; Proceedings of the IEEE/CVF Winter Conference on App
lications of Computer Vision (WACV), 2023, pp. 5252-5261
Due to the trending need for autonomous driving, camera-based object detection h
as recently attracted lots of attention and successful development. However, the
re are times when unexpected and severe weather occurs in outdoor environments,
making the detection tasks less effective and unexpected. In this case, addition
al sensors like lidar and radar are adopted to help the camera work in bad weath
er. However, existing multimodal detection methods do not consider the character
istics of different vehicle sensors to complement each other. Therefore, a novel
 end-to-end multimodal multistage object detection network called MT-DETR is pro

posed. Unlike the unimodal object detection networks, MT-DETR adds fusion modules and enhancement modules and adopts a hierarchical fusion mechanism. The Residual Fusion Module (RFM) and Confidence Fusion Module (CFM) are designed to fuse camera, lidar, radar, and time features. The Residual Enhancement Module (REM) reinforces each unimodal branch while a multistage loss is introduced to strengthen each branch's effectiveness. The synthesis algorithm for generating camera-lidar data pairs in foggy conditions further boosts the performance in unseen adverse weather. Extensive experiments on various weather conditions of the STF dataset demonstrate that MT-DETR outperforms state-of-the-art methods. The generality of MT-DETR has also been confirmed by replacing the feature extractor in the experiments. The code and pre-trained models are available on https://github.com/Chushihyun/MT-DETR.

**********

Video Object Matting via Hierarchical Space-Time Semantic Guidance

Yumeng Wang, Bo Xu, Ziwen Li, Han Huang, Cheng Lu, Yandong Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5120-5129

Different from most existing approaches that require trimap generation for each frame, we reformulate video object matting (VOM) by introducing improved semantic guidance propagation. The proposed approach can achieve a higher degree of temporal coherence between frames with only a single coarse mask as reference. In this paper, we adapt the hierarchical memory matching mechanism into the space-time baseline to build an efficient and robust framework for semantic guidance propagation and alpha prediction. To enhance the temporal smoothness, we also propose a cross-frame attention refinement (CFAR) module that can refine the feature representations across multiple adjacent frames (both historical and current frames) based on the spatio-temporal correlation among the cross-frame pixels. Extensive experiments demonstrate the effectiveness of hierarchical spatio-temporal semantic guidance and the cross-video-frame attention refinement module, and our model outperforms the state-of-the-art VOM methods. We also analyze the significance of different components in our model.

**********

PatchDropout: Economizing Vision Transformers Using Patch Dropout

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, Kevin Smith; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3953-3962

Vision transformers have demonstrated the potential to outperform CNNs in a variety of vision tasks. But the computational and memory requirements of these models prohibit their use in many applications, especially those that depend on high-resolution images, such as medical image classification. Efforts to train ViTs more efficiently are overly complicated, necessitating architectural changes or intricate training schemes. In this work, we show that standard ViT models can be efficiently trained at high resolution by randomly dropping input image patches. This simple approach, PatchDropout, reduces FLOPs and memory by at least 50% in standard natural image datasets such as ImageNet, and those savings only increase with image size. On CSAW, a high-resolution medical dataset, we observe a 5x savings in computation and memory using PatchDropout, along with a boost in performance. For practitioners with a fixed computational or memory budget, PatchDropout makes it possible to choose image resolution, hyperparameters, or model size to get the most performance out of their model.

**********

Reducing Annotation Effort by Identifying and Labeling Contextually Diverse Classes for Semantic Segmentation Under Domain Shift

Sharat Agarwal, Saket Anand, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5904-5913

In Active Domain Adaptation (ADA), one uses Active Learning (AL) to select target domain frames to annotate for Domain Adaptation (DA). Thus, ADA creates a continuum of cost-performance trade-off models, with unsupervised, and fully supervised DA techniques at the two ends. We observe that in ADA not all regions of a selected frame contribute equally to a model's performance, and there is a strong

correlation between annotating certain hard/unique/novel object/stuff instances, and a model's performance. E.g., road regions in a target dataset may look mostly similar to source domain except for certain curved instances, where annotation may be more useful. Based on the observation, we propose Anchor-based and Augmentation-based ADA techniques, which, given a selected frame, determine certain 'hard' semantic regions to be annotated in that frame, such that the selected regions are complementary and diverse in the context of the current labeled set. The proposed techniques carefully avoid the pitfall of region based AL techniques which try to choose most uncertain regions in a frame, but ends up selecting all edge pixels, and similar annotation cost as the whole frame. We show that our approach achieves 66.6 \miou on \gta->\cityscapes dataset with a budget of 4.7% in comparison to 64.9 \miou by MADA [??]. Our technique can also be used as a decorator for any existing frame-based AL technique. E.g., we report 1.5% performance improvement for CDAL [??] on \cityscapes using our approach.
**********

Color Recommendation for Vector Graphic Documents Based on Multi-Palette Representation

Qianru Qiu, Xueting Wang, Mayu Otani, Yuki Iwazaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3621-3629

Vector graphic documents present multiple visual elements, such as images, shapes, and texts. Choosing appropriate colors for multiple visual elements is a difficult but crucial task for both amateurs and professional designers. Instead of creating a single color palette for all elements, we extract multiple color palettes from each visual element in a graphic document, and then combine them into a color sequence. We propose a masked color model for color sequence completion and recommend the specified colors based on color context in multi-palette with high probability. We train the model and build a color recommendation system on a large-scale dataset of vector graphic documents. The proposed color recommendation method outperformed other state-of-the-art methods by both quantitative and qualitative evaluations on color prediction and our color recommendation system received positive feedback from professional designers in an interview study.
**********

Interactive Image Manipulation With Complex Text Instructions

Ryugo Morita, Zhiqiang Zhang, Man M. Ho, Jinjia Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1053-1062

Recently, text-guided image manipulation has received increasing attention in the research field of multimedia processing and computer vision due to its high flexibility and controllability. Its goal is to semantically manipulate parts of an input reference image according to the text descriptions. However, most of the existing works have the following problems: (1) text-irrelevant content cannot always be maintained but randomly changed, (2) the performance of image manipulation still needs to be further improved, (3) only can manipulate descriptive attributes. To solve these problems, we propose a novel image manipulation method that interactively edits an image using complex text instructions. It allows users to not only improve the accuracy of image manipulation but also achieve complex tasks such as enlarging, dwindling, or removing objects and replacing the background with the input image. To make these tasks possible, we apply three strategies. First, the given image is divided into text-relevant content and text-irrelevant content. Only the text-relevant content is manipulated and the text-irrelevant content can be maintained. Second, a super-resolution method is used to enlarge the manipulation region to further improve the operability and to help manipulate the object itself. Third, a user interface is introduced for editing the segmentation map interactively to re-modify the generated image according to the user's desires. Extensive experiments on the Caltech-UCSD Birds-200-2011 (CUB) dataset and Microsoft Common Objects in Context (MS COCO) datasets demonstrate our proposed method can enable interactive, flexible, and accurate image manipulation in real-time. Through qualitative and quantitative evaluations, we show that the proposed model outperforms other state-of-the-art methods.
**********

DSTrans: Dual-Stream Transformer for Hyperspectral Image Restoration
Dabing Yu, Qingwu Li, Xiaolin Wang, Zhiliang Zhang, Yixi Qian, Chang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3739-3749
Most CNN models exhibit two major flaws in hyperspectral image (HSI) restoration tasks. First, limited high-dimensional HSI training examples exacerbate the difficulty of deep learning methods in learning effective spatial and spectral representations. Second, the existing CNN-based methods model local relations and present limitations in capturing long-range dependencies. In this paper, we customize a novel dual-stream Transformer (DSTrans) for HSI restoration, which mainly consists of the dual-stream attention and the dual-stream feed-forward network. Specifically, we develop the dual-stream attention consisting of Multi-Dconv-head spectral attention (MDSA) and Multi-head Spatial self-attention (MSSA). MDSA and MSSA respectively calculate self-attention along the spectral and spatial dimensions in local windows to capture long-range spectrum dependencies and model global spatial interactions. Meanwhile, the dual-stream feed-forward network is developed to extract global signals and local details in parallel branches. In addition, we exploit a multi-tasking network to train the auxiliary RGB image (RGBI) task and HSI task jointly so that both numerous RGBI samples and limited HSI samples are exploited to learn parameter distribution for DSTrans. Extensive experimental results demonstrate that our method achieves state-of-the-art results on HSI restoration tasks, including HSI super-resolution and denoising. The source code can be obtained at: https://github.com/yudadabing/Dual-Stream-Transformer-for-Hyperspectral-Image-Restoration.
**********

LineEX: Data Extraction From Scientific Line Charts
Shivasankaran V. P., Muhammad Yusuf Hassan, Mayank Singh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6213-6221
In this paper, we introduce LineEX that extracts data from scientific line charts. We adapt existing vision transformers and pose detection methods and showcase significant performance gains over existing SOTA baselines. We also propose a new loss function and present its effectiveness against existing loss functions. In addition, we synthetically created the largest line chart dataset comprising 430K images. The code and the dataset will be placed in the public domain soon after the acceptance.
**********

Neural Implicit Representations for Physical Parameter Inference From a Single Video
Florian Hofherr, Lukas Koestler, Florian Bernard, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2093-2103
Neural networks have recently been used to analyze diverse physical systems and to identify the underlying dynamics. While existing methods achieve impressive results, they are limited by their strong demand for training data and their weak generalization abilities to out-of-distribution data. To overcome these limitations, we propose to combine neural implicit representations for appearance modeling with neural ordinary differential equations (ODEs) for modelling physical phenomena to obtain a dynamic scene representation that can be identified directly from visual observations. Our proposed model combines several unique advantages: (i) Contrary to existing approaches that require large training datasets, we are able to identify physical parameters from only a single video. (ii) The use of neural implicit representations enables the processing of high-resolution videos and the synthesis of photo-realistic images. (iii) The embedded neural ODE has a known parametric form that allows for the identification of interpretable physical parameters, and (iv) long-term prediction in state space. (v) Furthermore, the photo-realistic rendering of novel scenes with modified physical parameters becomes possible.
**********

Mesh-Tension Driven Expression-Based Wrinkles for Synthetic Faces

Chirag Raman, Charlie Hewitt, Erroll Wood, Tadas Baltrušaitis; Proceedings of th e IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp . 3515-3525
Recent advances in synthesizing realistic faces have shown that synthetic traini ng data can replace real data for various face-related computer vision tasks. A question arises: how important is realism? Is the pursuit of photorealism excess ive? In this work, we show otherwise. We boost the realism of our synthetic face s by introducing dynamic skin wrinkles in response to facial expressions, and ob serve significant performance improvements in downstream computer vision tasks. Previous approaches for producing such wrinkles either required prohibitive arti st effort to scale across identities and expressions, or were not capable of rec onstructing high-frequency skin details with sufficient fidelity. Our key contri bution is an approach that produces realistic wrinkles across a large and divers e population of digital humans. Concretely, we formalize the concept of mesh-ten sion and use it to aggregate possible wrinkles from high-quality expression scan s into albedo and displacement texture maps. At synthesis, we use these maps to produce wrinkles even for expressions not represented in the source scans. Addit ionally, to provide a more nuanced indicator of model performance under deformat ions resulting from compressed expressions, we introduce the 300W-winks evaluati on subset and the Pexels dataset of closed eyes and winks.
**********

CRT-6D: Fast 6D Object Pose Estimation With Cascaded Refinement Transformers
Pedro Castro, Tae-Kyun Kim; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 5746-5755
Learning based 6D object pose estimation methods rely on computing large interme diate pose representations and/or iteratively refining an initial estimation wit h a slow render-compare pipeline. This paper introduces a novel method we call C ascaded Pose Refinement Transformers, or CRT-6D. We replace the commonly used de nse intermediate representation with a sparse set of features sampled from the f eature pyramid we call OSKFs(Object Surface Keypoint Features) where each elemen t corresponds to an object keypoint. We employ lightweight deformable transforme rs and chain them together to iteratively refine proposed poses over the sampled OSKFs. We achieve inference runtimes 2x faster than the closest real-time state of the art methods while supporting up to 21 objects on a single model. We demo nstrate the effectiveness of CRT-6D by performing extensive experiments on the L M-O and YCBV datasets. Compared to real-time methods, we achieve state of the ar t on LM-O and YCB-V, falling slightly behind methods with inference runtimes one order of magnitude higher. The source code is available at: https://github.com/ PedroCastro/CRT-6D
**********

DCVNet: Dilated Cost Volume Networks for Fast Optical Flow
Huaizu Jiang, Erik Learned-Miller; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5150-5157
The cost volume, capturing the similarity of possible correspondences across two input images, is a key ingredient in state-of-the-art optical flow approaches. When sampling correspondences to build the cost volume, a large neighborhood rad ius is required to deal with large displacements, introducing a significant comp utational burden. To address this, coarse-to-fine or recurrent processing of the cost volume is usually adopted, where correspondence sampling in a local neighb orhood with a small radius suffices. In this paper, we propose an alternative by constructing cost volumes with different dilation factors to capture small and large displacements simultaneously. A U-Net with sikp connections is employed to convert the dilated cost volumes into interpolation weights between all possibl e captured displacements to get the optical flow. Our proposed model DCVNet only needs to process the cost volume once in a simple feedforward manner and does n ot rely on the sequential processing strategy. DCVNet obtains comparable accurac y to existing approaches and achieves real-time inference (30 fps on a mid-end 1 080ti GPU).
**********

Out-of-Distribution Detection via Frequency-Regularized Generative Models

Mu Cai, Yixuan Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5521-5530

Modern deep generative models can assign high likelihood to inputs drawn from outside the training distribution, posing threats to models in open-world deployments. While much research attention has been placed on defining new test-time measures of OOD uncertainty, these methods do not fundamentally change how deep generative models are regularized and optimized in training. In particular, generative models are shown to overly rely on the background information to estimate the likelihood. To address the issue, we propose a novel frequency-regularized learning (FRL) framework for OOD detection, which incorporates high-frequency information into training and guides the model to focus on semantically relevant features. FRL effectively improves performance on a wide range of generative architectures, including variational auto-encoder, GLOW, and PixelCNN++. On a new large-scale evaluation task, FRL achieves the state-of-the-art performance, outperforming a strong baseline Likelihood Regret by 10.7% (AUROC) while achieving 147x faster inference speed. Extensive ablations show that FRL improves the OOD detection performance while preserving the image generation quality.

**********

Similarity Contrastive Estimation for Self-Supervised Soft Contrastive Learning

Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault, Stéphane Canu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2706-2716

Contrastive representation learning has proven to be an effective self-supervised learning method. Most successful approaches are based on Noise Contrastive Estimation (NCE) and use different views of an instance as positives that should be contrasted with other instances, called negatives, that are considered as noise. However, several instances in a dataset are drawn from the same distribution and share underlying semantic information. A good data representation should contain relations, or semantic similarity, between the instances. Contrastive learning implicitly learns relations but considering all negatives as noise harms the quality of the learned relations. To circumvent this issue, we propose a novel formulation of contrastive learning using semantic similarity between instances called Similarity Contrastive Estimation (SCE). Our training objective is a soft contrastive learning one. Instead of hard classifying positives and negatives, we estimate from one view of a batch a continuous distribution to push or pull instances based on their semantic similarities. This target similarity distribution is sharpened to eliminate noisy relations. The model predicts for each instance, from another view, the target distribution while contrasting its positive with negatives. Experimental results show that SCE is Top-1 on the ImageNet linear evaluation protocol at 100 pretraining epochs with 72.1% accuracy and is competitive with state-of-the-art algorithms by reaching 75.4% for 200 epochs with multi-crop. We also show that SCE is able to generalize to several tasks. Source code is available here: https://github.com/CEA-LIST/SCE.

**********

HyperPosePDF - Hypernetworks Predicting the Probability Distribution on SO(3)

Timon Höfer, Benjamin Kiefer, Martin Messmer, Andreas Zell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2369-2379

Pose estimation of objects in images is an essential problem in virtual and augmented reality and robotics. Traditional solutions use depth cameras, which are expensive, and working solutions require long processing times. This work focuses on the more difficult task when only RGB information is available. To this end, we predict not only the pose of an object but the complete probability density function (pdf) on the rotation manifold. This is the most general way to approach the pose estimation problem and is particularly useful in analysing object symmetries. In this work, we leverage implicit neural representations for the task of pose estimation and show that hypernetworks can be used to predict the rotational pdf. Furthermore, we analyse the Fourier embedding on SO(3) and evaluate the effectiveness of an initial Fourier embedding that proved successful. Our HyperPosePDF outperforms the current SOTA approach on the SYMSOL dataset.

**********
Class-Level Confidence Based 3D Semi-Supervised Learning
Zhimin Chen, Longlong Jing, Liang Yang, Yingwei Li, Bing Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 633-642
Current pseudo-labeling strategies in 3D semi-supervised learning (SSL) fail to dynamically incorporate the variance of learning status which is affected by each class's learning difficulty and data imbalance. To address this problem, we practically demonstrate that 3D unlabeled data class-level confidence can represent the learning status. Based on this finding, we present a novel class-level confidence based 3D SSL method. Firstly, a dynamic thresholding strategy is proposed to utilize more unlabeled data, especially for low learning status classes. Then, a re-sampling strategy is designed to avoid biasing toward high learning status classes, which dynamically changes the sampling probability of each class. Unlike the latest state-of-the-art SSL method FlexMatch which also utilizes dynamic threshold, our method can be applied to the inherently imbalanced dataset and thus is more general. To show the effectiveness of our method in 3D SSL tasks, we conduct extensive experiments on 3D SSL classification and detection tasks. Our method significantly outperforms state-of-the-art counterparts for both 3D SSL classification and detection tasks in all datasets.
**********
PIDS: Joint Point Interaction-Dimension Search for 3D Point Cloud
Tunhou Zhang, Mingyuan Ma, Feng Yan, Hai Li, Yiran Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1298-1307
The interaction and dimension of points are two important axes in designing point operators to serve hierarchical 3D models. Yet, these two axes are heterogeneous and challenging to fully explore. Existing works craft point operator under a single axis and reuse the crafted operator in all parts of 3D models. This overlooks the opportunity to better combine point interactions and dimensions by exploiting varying geometry/density of 3D point clouds. In this work, we establish PIDS, a novel paradigm to jointly explore point interactions and point dimensions to serve semantic segmentation on point cloud data. We establish a large search space to jointly consider versatile point interactions and point dimensions. This supports point operators with various geometry/density considerations. The enlarged search space with heterogeneous search components calls for a better ranking of candidate models. To achieve this, we improve the search space exploration by leveraging predictor-based Neural Architecture Search (NAS), and enhance the quality of prediction by assigning unique encoding to heterogeneous search components based on their priors. We thoroughly evaluate the networks crafted by PIDS on two semantic segmentation benchmarks, showing  1% mIOU improvement on SemanticKITTI and S3DIS over state-of-the-art 3D models.
**********
Adaptive Feature Fusion for Cooperative Perception Using LiDAR Point Clouds
Donghao Qiao, Farhana Zulkernine; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1186-1195
Cooperative perception allows a Connected Autonomous Vehicle (CAV) to interact with the other CAVs in the vicinity to enhance perception of surrounding objects to increase safety and reliability. It can compensate for the limitations of the conventional vehicular perception such as blind spots, low resolution, and weather effects. An effective feature fusion model for the intermediate fusion methods of cooperative perception can improve feature selection and information aggregation to further enhance the perception accuracy. We propose adaptive feature fusion models with trainable feature selection modules. One of our proposed models Spatial-wise Adaptive feature Fusion (S-AdaFusion) outperforms all other State-of-the-Arts (SOTAs) on two subsets of the OPV2V dataset: Default CARLA Towns for vehicle detection and the Culver City for domain adaptation. In addition, previous studies have only tested cooperative perception for vehicle detection. A pedestrian, however, is much more likely to be seriously injured in a traffic accident. We evaluate the performance of cooperative perception for both vehicle and

pedestrian detection using the CODD dataset. Our architecture achieves higher Average Precision (AP) than other existing models for both vehicle and pedestrian detection on the CODD dataset. The experiments demonstrate that cooperative perception also improves the pedestrian detection accuracy compared to the conventional single vehicle perception process.

**********

HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar
Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, Jenq-Neng Hwang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5715-5724

This paper introduces a novel human pose estimation benchmark, Human Pose with Millimeter Wave Radar (HuPR), that includes synchronized vision and radio signal components. This dataset is created using cross-calibrated mmWave radar sensors and a monocular RGB camera for cross-modality training of radar-based human pose estimation. There are two advantages of using mmWave radar to perform human pose estimation. First, it is robust to dark and low-light conditions. Second, it is not visually perceivable by humans and therefore, can be widely applied to applications with privacy concerns, e.g., surveillance systems in patient rooms. In addition to the benchmark, we propose a cross-modality training framework that leverages the ground-truth 2D keypoints representing human body joints for training, which are systematically generated from the pre-trained 2D pose estimation network based on a monocular camera input image, avoiding laborious manual label annotation efforts. The framework consists of a new radar pre-processing method that better extracts the velocity information from radar data, Cross- and Self-Attention Module (CSAM), to fuse multi-scale radar features, and Pose Refinement Graph Convolutional Networks (PRGCN), to refine the predicted keypoint confidence heatmaps. Our intensive experiments on the HuPR benchmark show that the proposed scheme achieves better human pose estimation performance with only radar data, as compared to traditional pre-processing solutions and previous radio-frequency-based methods. Our proposed scheme further outperforms state-of-the-art pointcloud-based methods.

**********

GeoFill: Reference-Based Image Inpainting With Better Geometric Understanding
Yunhan Zhao, Connelly Barnes, Yuqian Zhou, Eli Shechtman, Sohrab Amirghodsi, Charless Fowlkes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1776-1786

Reference-guided image inpainting restores image pixels by leveraging the content from another single reference image. The primary challenge is how to precisely place the pixels from the reference image into the hole region. Therefore, understanding the 3D geometry that relates pixels between two views is a crucial step towards building a better model. Given the complexity of handling various types of reference images, we focus on the scenario where the images are captured by freely moving the same camera around. Compared to the previous work, we propose a principled approach that does not make heuristic assumptions about the planarity of the scene. We leverage a monocular depth estimate and predict relative pose between cameras, then align the reference image to the target by a differentiable 3D reprojection and a joint optimization of relative pose and depth map scale and offset. Our approach achieves state-of-the-art performance on both RealEstate10K and MannequinChallenge dataset with large baselines, complex geometry and extreme camera motions. We experimentally verify our approach is also better at handling large holes.

**********

Boosting Vision Transformers for Image Retrieval
Chull Hwan Song, Jooyoung Yoon, Shunghyun Choi, Yannis Avrithis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 107-117

The explosive increase in vision transformers studies has shown remarkable progress in vision tasks such as image classification and detection. However, in instance-level image retrieval, transformers have not yet shown good performance compared to convolutional networks. We propose a number of improvements that make t

ransformers outperform the state of the art for the first time. (1) We show that a hybrid architecture is more effective than plain transformers, by a large margin. (2) We introduce two branches collecting global (classification token) and local (patch tokens) information, from which we form a global image epresentation. (3) In each branch, we collect multi-layer features from the transformer encoder, corresponding to skip connections across distant layers. (4) We enhance locality of interactions at the deeper layers of the encoder, which is the relative weakness of vision transformers. We train our model on all commonly used training sets and, for the first time, we make fair comparisons separately per training set. In all cases, we outperform previous models based on global representation.

**********

DELS-MVS: Deep Epipolar Line Search for Multi-View Stereo

Christian Sormann, Emanuele Santellani, Mattia Rossi, Andreas Kuhn, Friedrich Fraundorfer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3087-3096

We propose a novel approach for deep learning-based Multi-View Stereo (MVS). For each pixel in the reference image, our method leverages a deep architecture to search for the corresponding point in the source image directly along the corresponding epipolar line. We denote our method DELS-MVS: Deep Epipolar Line Search Multi-View Stereo. Previous works in deep MVS select a range of interest within the depth space, discretize it, and sample the epipolar line according to the resulting depth values: this can result in an uneven scanning of the epipolar line, hence of the image space. Instead, our method works directly on the epipolar line: this guarantees an even scanning of the image space and avoids both the need to select a depth range of interest, which is often not known a priori and can vary dramatically from scene to scene, and the need for a suitable discretization of the depth space. In fact, our search is iterative, which avoids the building of a cost volume, costly both to store and to process. Finally, our method performs a robust geometry-aware fusion of the estimated depth maps, leveraging a confidence predicted alongside each depth. We test DELS-MVS on the ETH3D, Tanks and Temples and DTU benchmarks and achieve competitive results with respect to state-of-the-art approaches.

**********

Uplift and Upsample: Efficient 3D Human Pose Estimation With Uplifting Transformers

Moritz Einfalt, Katja Ludwig, Rainer Lienhart; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2903-2913

The state-of-the-art for monocular 3D human pose estimation in videos is dominated by the paradigm of 2D-to-3D pose uplifting. While the uplifting methods themselves are rather efficient, the true computational complexity depends on the per-frame 2D pose estimation. In this paper, we present a Transformer-based pose uplifting scheme that can operate on temporally sparse 2D pose sequences but still produce temporally dense 3D pose estimates. We show how masked token modeling can be utilized for temporal upsampling within Transformer blocks. This allows to decouple the sampling rate of input 2D poses and the target frame rate of the video and drastically decreases the total computational complexity. Additionally, we explore the option of pre-training on large motion capture archives, which has been largely neglected so far. We evaluate our method on two popular benchmark datasets: Human3.6M and MPI-INF-3DHP. With an MPJPE of 45.0 mm and 46.9 mm, respectively, our proposed method can compete with the state-of-the-art while reducing inference time by a factor of 12. This enables real-time throughput with variable consumer hardware in stationary and mobile applications. We release our code and models at https://github.com/goldbricklemon/uplift-upsample-3dhpe

**********

Boosting Neural Video Codecs by Exploiting Hierarchical Redundancy

Reza Pourreza, Hoang Le, Amir Said, Guillaume Sautière, Auke Wiggers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5355-5364

In video compression, coding efficiency is improved by reusing pixels from previ

ously decoded frames via motion and residual compensation. We define two levels of hierarchical redundancy in video frames: 1) first-order: redundancy in pixel space, i.e, similarities in pixel values across neighboring frames, which is effectively captured using motion and residual compensation, 2) second-order: redundancy in motion and residual maps due to smooth motion in natural videos. While most of the existing neural video coding literature addresses first-order redundancy, we tackle the problem of capturing second-order redundancy in neural video codecs via predictors. We introduce generic motion and residual predictors that learn to extrapolate from previously decoded data. These predictors are lightweight, and can be employed with most neural video codecs in order to improve their rate-distortion performance. Moreover, while RGB is the dominant colorspace in neural video coding literature, we introduce general modifications for neural video codecs to embrace the YUV420 colorspace and report YUV420 results. Our experiments show that using our predictors with a well-known neural video codec leads to 38% and 34% bitrate saving in RGB and YUV420 colorspaces measured on the UVG dataset.

**********

GarSim: Particle Based Neural Garment Simulator

Lokender Tiwari, Brojeshwar Bhowmick; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4472-4481

We present a particle-based neural garment simulator (dubbed as GarSim) that can simulate template garments on the target arbitrary body poses. Existing learning-based methods majorly work for specific garment type (e.g. t-shirt, skirt, etc) or garment topology, and needs retraining for a new type of garment. Similarly, some methods focus on a particular fabric, body shape, and pose. To circumvent these limitations, our method fundamentally learns the physical dynamics of the garment vertices conditioned on underlying body shape, motion, and fabric properties to generalize across garment types, topology, and fabric along with different body shape and pose. In particular, we represent the garment as a graph, where the nodes represent the physical state of the garment vertices, and the edges represent the relation between the two nodes. The nodes and edges of the garment graph encode various properties of garments and the human body to compute the dynamics of the vertices through a learned message-passing. Learning of such dynamics of the garment vertices conditioned on underlying body motion and fabric properties enables our method to be trained simultaneously for multiple types of garments (e.g., tops, skirts, etc) with arbitrary mesh resolutions, varying topologies, and fabric properties. Our experimental results show that GarSim with less amount of training data not only outperforms the SOTA methods on challenging CLOTH3D dataset both qualitatively and quantitatively, but also works reliably well on the unseen poses obtained from YouTube videos, and give satisfactory results on unseen cloth types which were not present during the training.

**********

Event-Based RGB Sensing With Structured Light

Seyed Ehsan Marjani Bajestani, Giovanni Beltrame; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5458-5467

Event-based cameras (ECs) are bio-inspired sensors that asynchronously report pixel brightness changes. Due to their high dynamic range, pixel bandwidth, temporal resolution, low power consumption, and computational simplicity, they are beneficial for vision-based projects in challenging lighting conditions and they can detect fast movements with their microsecond response time. The first generation of ECs are monochrome, but color data is very useful and sometimes essential for certain vision-based applications. The latest technology enables manufacturers to build color ECs, trading off the size of the sensor and substantially reducing the resolution compared to monochrome models, despite having the same bandwidth. In addition, ECs only detect changes in light and do not show static or slowly moving objects. We introduce a method to detect full RGB events using a monochrome EC aided by a structured light projector. The projector emits rapidly changing RGB patterns of light beams on the scene, the reflection of which is captured by the EC. We combine the benefits of ECs and projection-based techniques and allow depth and color detection of static or moving objects with a commercial

TI LightCrafter 4500 projector and a monocular monochrome EC, paving the way for frameless RGB-D sensing applications. Our code is available publicly: github.com/MISTLab/event_based_rgbd_ros

**********

Saliency Guided Experience Packing for Replay in Continual Learning
Gobinda Saha, Kaushik Roy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5273-5283

Artificial learning systems aspire to mimic human intelligence by continually learning from a stream of tasks without forgetting past knowledge. One way to enable such learning is to store past experiences in the form of input examples in episodic memory and replay them when learning new tasks. However, performance of such method suffers as the size of the memory becomes smaller. In this paper, we propose a new approach for experience replay, where we select the past experiences by looking at the saliency maps which provide visual explanations for the model's decision. Guided by these saliency maps, we pack the memory with only the parts or patches of the input images important for the model's prediction. While learning a new task, we replay these memory patches with appropriate zero-padding to remind the model about its past decisions. We evaluate our algorithm on CIFAR-100, miniImageNet and CUB datasets and report better performance than the state-of-the-art approaches. With qualitative and quantitative analyses we show that our method captures richer summaries of past experiences without any memory increase, and hence performs well with small episodic memory.

**********

CTrGAN: Cycle Transformers GAN for Gait Transfer
Shahar Mahpod, Noam Gaash, Hay Hoffman, Gil Ben-Artzi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 371-381

We introduce a novel approach for gait transfer from unconstrained videos in-the-wild. In contrast to motion transfer, the objective here is not to imitate the source's motions by the target, but rather to replace the walking source with the target, while transferring the target's typical gait. Our approach can be trained only once with multiple sources and is able to transfer the gait of the target from unseen sources, eliminating the need for retraining for each new source independently. Furthermore, we propose a novel metrics for gait transfer based on gait recognition models that enable to quantify the quality of the transferred gait, and show that existing techniques yield a discrepancy that can be easily detected.\nWe introduce Cycle Transformers GAN (CTrGAN), that consist of a decoder and encoder, both Transformers, where the attention is on the temporal domain between complete images rather than the spatial domain between patches. Using a widely-used gait recognition dataset, we demonstrate that our approach is capable of producing over an order of magnitude more realistic personalized gaits than existing methods, even when used with sources that were not available during training. As part of our solution, we present a detector that determines whether a video is real or generated by our model.

**********

Handling Image and Label Resolution Mismatch in Remote Sensing
Scott Workman, Armin Hadzic, M. Usman Rafique; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3709-3718

Though semantic segmentation has been heavily explored in vision literature, unique challenges remain in the remote sensing domain. One such challenge is how to handle resolution mismatch between overhead imagery and ground-truth label sources, due to differences in ground sample distance. To illustrate this problem, we introduce a new dataset and use it to showcase weaknesses inherent in existing strategies that naively upsample the target label to match the image resolution. Instead, we present a method that is supervised using low-resolution labels (without upsampling), but takes advantage of an exemplar set of high-resolution labels to guide the learning process. Our method incorporates region aggregation, adversarial learning, and self-supervised pretraining to generate fine-grained predictions, without requiring high-resolution annotations. Extensive experiments demonstrate the real-world applicability of our approach.

**********
Aggregating Bilateral Attention for Few-Shot Instance Localization
He-Yen Hsieh, Ding-Jie Chen, Cheng-Wei Chang, Tyng-Luh Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6325-6334
Attention filtering under various learning scenarios has proven advantageous in enhancing the performance of many neural network architectures. The mainstream attention mechanism is established upon the non-local block, also known as an essential component of the prominent Transformer networks, to catch long-range correlations. However, such unilateral attention is often hampered by sparse and obscure responses, revealing insufficient dependencies across images/patches, and high computational cost, especially for those employing the multi-head design. To overcome these issues, we introduce a novel mechanism of aggregating bilateral attention (ABA) and validate its usefulness in tackling the task of few-shot instance localization, reflecting the underlying query-support dependency. Specifically, our method facilitates uncovering informative features via assessing: i) an embedding norm for exploring the semantically-related cues; ii) context awareness for correlating the query data and support regions. ABA is then carried out by integrating the affinity relations derived from the two measurements to serve as a lightweight but effective query-support attention mechanism with high localization recall. We evaluate ABA on two localization tasks, namely, few-shot action localization and one-shot object detection. Extensive experiments demonstrate that the proposed ABA achieves superior performances over existing methods.
**********
AnoLeaf: Unsupervised Leaf Disease Segmentation via Structurally Robust Generative Inpainting
Swati Bhugra, Vinay Kaushik, Amit Gupta, Brejesh Lall, Santanu Chaudhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6415-6424
Plant diseases severely limits agriculture production, necessitating the high-throughput monitoring of plant leaves. Currently, this is formulated as an automatic disease segmentation task addressed via deep learning frameworks. These deep leaning frameworks trained with leaf image data in a supervised paradigm have few limitations, mainly: (1) training datasets are heavily imbalanced towards healthy leaf images, (2) disease region annotation is labour-intensive and (3) due to the heterogeneity of disease symptoms, these frameworks lacks generalisability. In this paper, we reformulate disease segmentation as an anomaly localisation task. Specifically, we introduce a novel unsupervised framework (AnoLeaf) based on an edge-guided inpainting that optimises the learning of contextual attention on only healthy leaf images. The network utilisation on diseased leaf images results in reconstruction of its healthy counterparts, generating an inpainting error. The contextual attention maps reinforce the inpainting error to effectively localise the disease. Thus, AnoLeaf alleviates the acquisition and annotation of rare disease images. Additional experiments on MVTec anomaly detection dataset further demonstrate its generalisability.
**********
How To Practice VQA on a Resource-Limited Target Domain
Mingda Zhang, Rebecca Hwa, Adriana Kovashka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4451-4460
Visual question answering (VQA) is an active research area at the intersection of computer vision and natural language understanding. One major obstacle that keeps VQA models that perform well on benchmarks from being as successful on real-world applications, is the lack of annotated Image-Question-Answer triplets in the task of interest. In this work, we focus on a previously overlooked perspective, which is the disparate effectiveness of transfer learning and domain adaptation methods depending on the amount of labeled/unlabeled data available. We systematically investigated the visual domain gaps and question-defined textual gaps, and compared different knowledge transfer strategies under unsupervised, self-supervised, semi-supervised and fully-supervised adaptation scenarios. We show that different methods have varied sensitivity and requirements for data amount i

n the target domain. We conclude by sharing the best practice from our explorati on regarding transferring VQA models to resource-limited target domains.
**********

Lightweight Video Denoising Using Aggregated Shifted Window Attention
Lydia Lindner, Alexander Effland, Filip Ilic, Thomas Pock, Erich Kobler; Proceed ings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) , 2023, pp. 351-360
Video denoising is a fundamental problem in numerous computer vision application s. State-of-the-art attention-based denoising methods typically yield good resul ts, but require vast amounts of GPU memory and usually suffer from very long com putation times. Especially in the field of restoring digitized high-resolution h istoric films, these techniques are not applicable in practice. To overcome thes e issues, we introduce a lightweight video denoising network that combines effic ient axial-coronal-sagittal (ACS) convolutions with a novel shifted window atten tion formulation (ASwin), which is based on the memory-efficient aggregation of self- and cross-attention across video frames. We numerically validate the perfo rmance and efficiency of our approach on synthetic Gaussian noise. Moreover, we train our network as a general-purpose blind denoising model for real-world vide os, using a realistic noise synthesis pipeline to generate clean-noisy video pai rs. A user study and non- reference quality assessment prove that our method out performs the state-of-the-art on real-world historic videos in terms of denoisin g performance and temporal consistency.
**********

Dynamic Mixture of Counter Network for Location-Agnostic Crowd Counting
Mingjie Wang, Hao Cai, Yong Dai, Minglun Gong; Proceedings of the IEEE/CVF Winte r Conference on Applications of Computer Vision (WACV), 2023, pp. 167-177
Crowd counting has attracted increasing attentions in recent years due to its ch allenges and wide societal applications. Despite persevering efforts made by the research community, most of existing methods require a large amount of location -level annotations. Collecting such type of fine-granularity supervisory signals is extremely time-consuming and labour-intensive, thereby hindering the well ge neralization of these location-adherent models. To shun this drawback, several p ioneering studies open a promising research direction of location-agonistic crow d counting. Albeit the noticeable efforts, they somewhat ignore the merits of di verse learning paradigms and the issue of intractable density shift. To ameliora te these issues, in this paper, a novel Dynamic Mixture of Counter Network (DMCN et) is proposed for location-agnostic crowd counting. Specifically, our DMCNet i nherits the hybrid advantages of CNNs (e.g. locality-oriented and pyramidal prop erty) and MLP-based structure (e.g. global receptive fields and light weight). P articularly, the dynamic counter predictor and the mixture of counter heads are delicately designed to hammer at combating huge density shift and overfitting. E xtensive experiments demonstrate that our DMCNet attains state-of-the-art perfor mance against existing location-agnostic approaches and performs on par with man y conventional location-adherent ones.
**********

Fast Online Video Super-Resolution With Deformable Attention Pyramid
Dario Fuoli, Martin Danelljan, Radu Timofte, Luc Van Gool; Proceedings of the IE EE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 17 35-1744
Video super-resolution (VSR) has many applications that pose strict causal, real -time, and latency constraints, including video streaming and TV. We address the VSR problem under these settings, which poses additional important challenges s ince information from future frames is unavailable. Importantly, designing effic ient, yet effective frame alignment and fusion modules remain central problems. In this work, we propose a recurrent VSR architecture based on a deformable atte ntion pyramid (DAP). Our DAP aligns and integrates information from the recurren t state into the current frame prediction. To circumvent the computational cost of traditional attention-based methods, we only attend to a limited number of sp atial locations, which are dynamically predicted by the DAP. Comprehensive exper iments and analysis of the proposed key innovations show the effectiveness of ou

r approach. We significantly reduce processing time and computational complexity in comparison to state-of-the-art methods, while maintaining a high performance. We surpass state-of-the-art method EDVR-M on two standard benchmarks with a speed-up of over 3x.

**********

Perceiver-VL: Efficient Vision-and-Language Modeling With Iterative Latent Attention

Zineng Tang, Jaemin Cho, Jie Lei, Mohit Bansal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4410-4420

We present Perceiver-VL, a vision-and-language framework that efficiently handles high-dimensional multimodal inputs such as long videos and text. Powered by the iterative latent-cross-attention of Perceiver, our framework scales with linear complexity, in contrast to the quadratic complexity of self-attention used in many state-of-the-art transformer-based models. To further improve the efficiency of our framework, we also study applying LayerDrop on cross-attention layers and introduce a mixed-stream architecture for cross-modal retrieval. We evaluate Perceiver-VL on diverse video-text and image-text benchmarks, where Perceiver-VL achieves the lowest GFLOPs and latency, while maintaining competitive performance. In addition, we also provide comprehensive analyses over various aspects of our framework, including pretraining data, scalability of latent size and input size, dropping cross-attention layers at inference to reduce latency, modality aggregation strategy, positional encoding, and weight initialization strategy.

**********

Self-Supervised Pyramid Representation Learning for Multi-Label Visual Analysis and Beyond

Cheng-Yen Hsieh, Chih-Jung Chang, Fu-En Yang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2696-2705

While self-supervised learning has been shown to benefit a number of vision tasks, existing techniques mainly focus on image-level manipulation, which may not generalize well to downstream tasks at patch or pixel levels. Moreover, existing SSL methods might not sufficiently describe and associate the above representations within and across image scales. In this paper, we propose a Self-Supervised Pyramid Representation Learning (SS-PRL) framework. The proposed SS-PRL is designed to derive pyramid representations at patch levels via learning proper prototypes, with additional learners to observe and relate inherent semantic information within an image. In particular, we present a cross-scale patch-level correlation learning in SS-PRL, which allows the model to aggregate and associate information learned across patch scales. We show that, with our proposed SS-PRL for model pre-training, one can easily adapt and fine-tune the models for a variety of applications including multi-label classification, object detection, and instance segmentation.

**********

Nearest Neighbors Meet Deep Neural Networks for Point Cloud Analysis

Renrui Zhang, Liuhui Wang, Ziyu Guo, Jianbo Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1246-1255

Performances on standard 3D point cloud benchmarks have plateaued, resulting in oversized models and complex network design to make a fractional improvement. We present an alternative to enhance existing deep neural networks without any redesigning or extra parameters, termed as Spatial-Neighbor Adapter SN-Adapter. Building on any trained 3D network, we utilize its learned encoding capability to extract features of the training dataset and summarize them as prototypical spatial knowledge. For a test point cloud, the SN-Adapter retrieves k nearest neighbors (k-NN) from the pre-constructed spatial prototypes and linearly interpolates the k-NN prediction with that of the original 3D network. By providing complementary characteristics, the proposed SN-Adapter serves as a plug-and-play module to economically improve performance in a non-parametric manner. More importantly, our SN-Adapter can be effectively generalized to various 3D tasks, including shape classification, part segmentation, and 3D object detection, demonstrating its superiority and robustness. We hope our approach could show a new perspective

for point cloud analysis and facilitate future research.
**********

SIUNet: Sparsity Invariant U-Net for Edge-Aware Depth Completion
Avinash Nittur Ramesh, Fabio Giovanneschi, María A. González-Huici; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5818-5827
Depth completion is the task of generating dense depth images from sparse depth measurements, e.g., LiDARs. Existing unguided approaches fail to recover dense depth images with sharp object boundaries due to depth bleeding, especially from extremely sparse measurements. State-of-the-art guided approaches require additional processing for spatial and temporal alignment of multi-modal inputs, and sophisticated architectures for data fusion, making them non-trivial for customized sensor setup. To address these limitations, we propose an unguided approach based on UNet that is invariant to sparsity of inputs. Boundary consistency in reconstruction is explicitly enforced through auxiliary learning on a synthetic dataset with dense depth and depth contour images as targets, followed by fine-tuning on a real-world dataset. With our network architecture and simple implementation approach, we achieve competitive results among unguided approaches on KITTI benchmark and show that the reconstructed image has sharp boundaries and is robust even towards extremely sparse LiDAR measurements.
**********

Weakly Supervised Face Naming With Symmetry-Enhanced Contrastive Loss
Tingyu Qu, Tinne Tuytelaars, Marie-Francine Moens; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3505-3514
We revisit the weakly supervised cross-modal face-name alignment task; that is, given an image and a caption, we label the faces in the image with the names occurring in the caption. Whereas past approaches have learned the latent alignment between names and faces by uncertainty reasoning over a set of images and their respective captions, in this paper, we rely on appropriate loss functions to learn the alignments in a neural network setting and propose SECLA and SECLA-B. SECLA is a Symmetry-Enhanced Contrastive Learning-based Alignment model that can effectively maximize the similarity scores between corresponding faces and names in a weakly supervised fashion. A variation of the model, SECLA-B, learns to align names and faces as humans do, that is, learning from easy to hard cases to further increase the performance of SECLA. More specifically, SECLA-B applies a two-stage learning framework: (1) Training the model on an easy subset with a few names and faces in each image-caption pair. (2) Leveraging the known pairs of names and faces from the easy cases using a bootstrapping strategy with additional loss to prevent forgetting and learning new alignments at the same time. We achieve state-of-the-art results for both the augmented Labeled Faces in the Wild dataset and the Celebrity Together dataset. In addition, we believe that our methods can be adapted to other multimodal news understanding tasks.
**********

HoechstGAN: Virtual Lymphocyte Staining Using Generative Adversarial Networks
Georg Wölflein, In Hwa Um, David J. Harrison, Ognjen Arandjelovi■; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4997-5007
The presence and density of specific types of immune cells are important to understand a patient's immune response to cancer. However, immunofluorescence staining required to identify T cell subtypes is expensive, timeconsuming, and rarely performed in clinical settings. We present a framework to virtually stain Hoechst images (which are cheap and widespread) with both CD3 and CD8 to identify T cell subtypes in clear cell renal cell carcinoma using generative adversarial networks. Our proposed method jointly learns both staining tasks, incentivising the network to incorporate mutually beneficial information from each task. We devise a novel metric to quantify the virtual staining quality, and use it to evaluate our method.
**********

Switching to Discriminative Image Captioning by Relieving a Bottleneck of Reinforcement Learning

Ukyo Honda, Taro Watanabe, Yuji Matsumoto; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1124-1134
Discriminativeness is a desirable feature of image captions: captions should describe the characteristic details of input images. However, recent high-performing captioning models, which are trained with reinforcement learning (RL), tend to generate overly generic captions despite their high performance in various other criteria. First, we investigate the cause of the unexpectedly low discriminativeness and show that RL has a deeply rooted side effect of limiting the output words to high-frequency words. The limited vocabulary is a severe bottleneck for discriminativeness as it is difficult for a model to describe the details beyond its vocabulary. This identification of the bottleneck allows us to drastically recast discriminative image captioning as a much simpler task of encouraging low-frequency word generation. Hinted by long-tail classification and debiasing methods, we propose the methods that easily switch off-the-shelf RL models to discriminativeness-aware models with only a single-epoch fine-tuning on the part of the parameters. Extensive experiments demonstrate that our methods significantly enhance the discriminativeness of off-the-shelf RL models and even outperform previous discriminativeness-aware methods with much smaller computational costs. Detailed analysis and human evaluation also verify that our methods boost the discriminativeness without sacrificing the overall quality of captions.
**********

RANCER: Non-Axis Aligned Anisotropic Certification With Randomized Smoothing
Taras Rumezhak, Francisco Girbal Eiras, Philip H.S. Torr, Adel Bibi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4672-4680
As modern networks have been proven to be unprotected from adversarial attacks and are applied in safety-critical applications, defense against them is very crucial. Many works were dedicated to this topic, but randomized smoothing has been recently proven to be an effective approach for the certified defense of deep neural networks and getting robust classifiers. Some prior results were obtained utilizing the techniques of adding extra parameters to extend the limits of the certification regions. In this way, sample-wise optimization was proposed to maximize the certification radius per input. The idea was further extended with the generalized anisotropic counterparts of l1 and l2 certificates which allow achieving larger certified region volume avoiding worst-case certification near potentially larger safe regions. However, anisotropic certification is limited by the aligned axis lacking the freedom to extend in any direction. To mitigate this constraint, in this work, we (i) revisit the anisotropic certification, provide an analysis of its non-axis aligned counterpart and propose its rotation-free extension, (ii) conduct experiments on the CIFAR-10 dataset to report the improved performance.
**********

HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation
Yintong Wang, LiLi Chen, Jiamao Li, Xiaolin Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5675-5684
Despite the substantial progress in 3D hand pose estimation, inferring plausible and accurate poses in the presence of severe self-occlusion and high self-similarity remains an inherent challenge. To mitigate the ambiguity arising from invisible and similar joints, we propose a novel Topology-aware Transformer network named HandGCNFormer, incorporating the prior knowledge of hand kinematic topology into the network while modeling long-range context information. Specifically, we present a novel Graphformer decoder with an additional node-offset graph convolutional layer (NoffGConv) that optimizes the synergy of Transformer and GCN, capturing long-range dependencies as well as local topology connection between joints. Furthermore, we replace the standard MLP prediction head with a novel Topology-aware head to better utilize local topology constraints for more plausible and accurate poses. Our method achieves state-of-the-art performance on four challenging datasets including Hands2017, NYU, ICVL, and MSRA.
**********

Mixture Outlier Exposure: Towards Out-of-Distribution Detection in Fine-Grained Environments
Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, Hai Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5531-5540
Many real-world scenarios in which DNN-based recognition systems are deployed have inherently fine-grained attributes (e.g., bird-species recognition, medical image classification). In addition to achieving reliable accuracy, a critical subtask for these models is to detect Out-of-distribution (OOD) inputs. Given the nature of the deployment environment, one may expect such OOD inputs to also be fine-grained w.r.t. the known classes (e.g., a novel bird species), which are thus extremely difficult to identify. Unfortunately, OOD detection in fine-grained scenarios remains largely underexplored. In this work, we aim to fill this gap by first carefully constructing four large-scale fine-grained test environments, in which existing methods are shown to have difficulties. Particularly, we find that even explicitly incorporating a diverse set of auxiliary outlier data during training does not provide sufficient coverage over the broad region where fine-grained OOD samples locate. We then propose Mixture Outlier Exposure (MixOE), which mixes ID data and training outliers to expand the coverage of different OOD granularities, and trains the model such that the prediction confidence linearly decays as the input transitions from ID to OOD. Extensive experiments and analyses demonstrate the effectiveness of MixOE for building up OOD detector in fine-grained environments. The code is available at https://github.com/zjysteven/MixOE.
**********
MixVPR: Feature Mixing for Visual Place Recognition
Amar Ali-bey, Brahim Chaib-draa, Philippe Giguère; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2998-3007
Visual Place Recognition (VPR) is a crucial part of mobile robotics and autonomous driving as well as other computer vision tasks. It refers to the process of identifying a place depicted in a query image using only computer vision. At large scale, repetitive structures, weather, and illumination changes pose a real challenge, as appearances can drastically change over time. Along with tackling these challenges, an efficient VPR technique must also be practical in real-world scenarios where latency matters. To address this, we introduce MixVPR, a new holistic feature aggregation technique that takes feature maps from pre-trained backbones as a set of global features. Then, it incorporates a global relationship between elements in each feature map in a cascade of feature mixing, eliminating the need for local or pyramidal aggregation as done in NetVLAD or TransVPR. We demonstrate the effectiveness of our technique through extensive experiments on multiple large-scale benchmarks. Our method outperforms all existing techniques by a large margin while having less than half the number of parameters compared to CosPlace and NetVLAD. We achieve a new all-time high recall@1 score of 94.6% on Pitts250k-test, 88.0% on MapillarySLS, and more importantly, 58.4% on Nordland. Finally, our method outperforms two-stage retrieval techniques such as Patch-NetVLAD, TransVPR and SuperGLUE, all while being orders of magnitude faster.
**********
Uncertainty-Aware Interactive LiDAR Sampling for Deep Depth Completion
Kensuke Taguchi, Shogo Morita, Yusuke Hayashi, Wataru Imaeda, Hironobu Fujiyoshi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3028-3036
Programmable scan LiDAR is able to measure arbitrary areas and is expected to be used in various applications. In this paper, we study a LiDAR sampling strategy for deep depth completion of a programmable scan LiDAR with an RGB camera. General data sampling strategies include adaptive approaches such as active learning, in which candidate data are assessed through a task model for data selection and then the selected data pool is updated sequentially. Although it is an effective approach, the adaptive approach requires many iterations involving the inference process to assess the candidate data, which is not suitable for LiDAR systems. Therefore, we propose a novel interactive LiDAR sampling method without each

inference process. Our key insights are that we assess sampling candidates by depth estimation uncertainty and virtually update the uncertainty by an approximation of the candidate assessment. This enables us to add interactivity to the model state without requiring each inference process. We demonstrate the effectiveness of our method on the KITTI dataset and the generalization performance on the NYU-Depth-v2 dataset in comparison with a conventional adaptive LiDAR sampling method, and we find superior results in the depth completion task. We also show ablation studies to analyze our approach.

**********

Contrastive Knowledge-Augmented Meta-Learning for Few-Shot Classification

Rakshith Subramanyam, Mark Heimann, T.S. Jayram, Rushil Anirudh, Jayaraman J. Thiagarajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2479-2487

Model agnostic meta-learning algorithms aim to infer priors from several observed tasks that can then be used to adapt to a new task with few examples. Given the inherent diversity of tasks arising in existing benchmarks, recent methods have resorted to task-specific adaptation of the prior. Our goal is to improve generalization of meta learners when the task distribution contains challenging distribution shifts and semantic disparities. To this end, we introduce CAML (Contrastive Knowledge-Augmented Meta Learning), a knowledge-enhanced few-shot learning approach that evolves a knowledge graph to encode historical experience, and employs a contrastive distillation strategy to leverage the encoded knowledge for task-aware modulation of the base learner. In addition to the standard few-shot task adaptation, we also consider the more challenging multi-domain task adaptation and few-shot dataset generalization settings in our evaluation with standard benchmarks. Our empirical study shows that CAML (i) enables simple task encoding schemes; (ii) eliminates the need for knowledge extraction at inference time; and most importantly, (iii) effectively aggregates historical experience thus leading to improved performance in both multi-domain adaptation and dataset generalization.

**********

RSF: Optimizing Rigid Scene Flow From 3D Point Clouds Without Labels

David Deng, Avideh Zakhor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1277-1286

We present a method for optimizing object-level rigid 3D scene flow over two successive point clouds without any annotated labels in autonomous driving settings. Rather than using pointwise flow vectors, our approach represents scene flow as the composition a global ego-motion and a set of bounding boxes with their own rigid motions, exploiting the multi-body rigidity commonly present in dynamic scenes. We jointly optimize these parameters over a novel loss function based on the nearest neighbor distance using a differentiable bounding box formulation. Our approach achieves state-of-the-art accuracy on KITTI Scene Flow and nuScenes without requiring any annotations, outperforming even supervised methods. Additionally, we demonstrate the effectiveness of our approach on motion segmentation and ego-motion estimation. Lastly, we visualize our predictions and validate our loss function design with an ablation study.

**********

Semi-Supervised Learning for Low-Light Image Restoration Through Quality Assisted Pseudo-Labeling

Sameer Malik, Rajiv Soundararajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4105-4114

Convolutional neural networks have been successful in restoring images captured under poor illumination conditions addressing multiple challenges such as contrast enhancement, denoising, and color cast removal. Nevertheless, such approaches require a large number of paired low-light and ground truth images for training. Thus, we study the problem of semi-supervised learning for low-light image restoration when limited low-light images have ground truth labels. Our main contributions in this work are twofold. We first deploy an ensemble of low-light restoration networks to restore the unlabeled images and generate a set of potential pseudo-labels. We model the contrast distortions in the labeled set to generate

different sets of training data and create the ensemble of networks. We then design a contrastive self-supervised learning based image quality measure to obtain the pseudo-label among the images restored by the ensemble. We show that training the restoration network with the pseudo-labels allows us to achieve excellent restoration performance even with very few labeled pairs. We conduct extensive experiments on three popular low-light image restoration datasets to show the superior performance of our semi-supervised low-light image restoration compared to other approaches.
**********

Select, Label, and Mix: Learning Discriminative Invariant Feature Representations for Partial Domain Adaptation
Aadarsh Sahoo, Rameswar Panda, Rogerio Feris, Kate Saenko, Abir Das; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4210-4219
Partial domain adaptation which assumes that the unknown target label space is a subset of the source label space has attracted much attention in computer vision. Despite recent progress, existing methods often suffer from three key problems: negative transfer, lack of discriminability, and domain invariance in the latent space. To alleviate the above issues, we develop a novel 'Select, Label, and Mix' (SLM) framework that aims to learn discriminative invariant feature representations for partial domain adaptation. First, we present an efficient "select" module that automatically filters out the outlier source samples to avoid negative transfer while aligning distributions across both domains. Second, the "label" module iteratively trains the classifier using both the labeled source domain data and the generated pseudo-labels for the target domain to enhance the discriminability of the latent space. Finally, the "mix" module utilizes domain mixup regularization jointly with the other two modules to explore more intrinsic structures across domains leading to a domain-invariant latent space for partial domain adaptation. Extensive experiments on several benchmark datasets demonstrate the superiority of our proposed framework over state-of-the-art methods. Project page: https://cvir.github.io/projects/slm.
**********

Weakly-Supervised Optical Flow Estimation for Time-of-Flight
Michael Schelling, Pedro Hermosilla, Timo Ropinski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2135-2144
Indirect Time-of-Flight (iToF) cameras are a widespread type of 3D sensor, which perform multiple captures to obtain depth values of the captured scene. While recent approaches to correct iToF depths achieve high performance when removing multi-path-interference and sensor noise, little research has been done to tackle motion artifacts. In this work we propose a training algorithm, which allows to supervise Optical Flow (OF) networks directly on the reconstructed depth, without the need of having ground truth flows. We demonstrate that this approach enables the training of OF networks to align raw iToF measurements and compensate motion artifacts in the iToF depth images. The approach is evaluated for both single- and multi-frequency sensors as well as multi-tap sensors, and is able to outperform other motion compensation techniques.
**********

THOR-Net: End-to-End Graformer-Based Realistic Two Hands and Object Reconstruction With Self-Supervision
Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1001-1010
Realistic reconstruction of two hands interacting with objects is a new and challenging problem that is essential for building personalized Virtual and Augmented Reality environments. Graph Convolutional networks (GCNs) allow for the preservation of the topologies of hands poses and shapes by modeling them as a graph. In this work, we propose the THOR-Net which combines the power of GCNs, Transformer, and self-supervision to realistically reconstruct two hands and an object from a single RGB image. Our network comprises two stages; namely the features extraction stage and the reconstruction stage. In the features extraction stage, a

Keypoint RCNN is used to extract 2D poses, features maps, heatmaps, and boundin g boxes from a monocular RGB image. Thereafter, this 2D information is modeled a s two graphs and passed to the two branches of the reconstruction stage. The sha pe reconstruction branch estimates meshes of two hands and an object using our n ovel coarse-to-fine GraFormer shape network. The 3D poses of the hands and objec ts are reconstructed by the other branch using a GraFormer network. Finally, a s elf-supervised photometric loss is used to directly regress the realistic textur ed of each vertex in the hands' meshes. Our approach achieves State-of-the-art r esults in Hand shape estimation on the HO3D dataset (10.0mm) exceeding ArtiBoost (10.8mm). It also surpasses other methods in hand pose estimation on the challe nging two hands and object (H2O) dataset by 5mm on the left-hand pose and 1 mm o n the right-hand pose. The code base of THOR-Net will be released soon under htt ps://github.com/ATAboukhadra/THOR-Net.

**********

Discrete Cosin TransFormer: Image Modeling From Frequency Domain

Xinyu Li, Yanyi Zhang, Jianbo Yuan, Hanlin Lu, Yibo Zhu; Proceedings of the IEEE /CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5468 -5478

In this paper, we propose Discrete Cosin TransFormer (DCFormer) that directly le arn semantics from DCT-based frequency domain representation. We first show that transformer-based networks are able to learn semantics directly from frequency domain representation based on discrete cosine transform (DCT) without compromis ing the performance. To achieve the desired efficiency-effectiveness trade-off, we then leverage an input information compression on its frequency domain repres entation, which highlights the visually significant signals inspired by JPEG com pression. We explore different frequency domain down-sampling strategies and sho w that it is possible to preserve the semantic meaningful information by strateg ically dropping the high-frequency components. The proposed DCFormer is tested o n various downstream tasks including image classification, object detection and instance segmentation, and achieves state-of-the-art comparable performance with less FLOPs, and outperforms the commonly used backbone (e.g. SWIN) at similar F LOPs. Our ablation results also show that the proposed method generalizes well o n different transformer backbones.

**********

Intra-Source Style Augmentation for Improved Domain Generalization

Yumeng Li, Dan Zhang, Margret Keuper, Anna Khoreva; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 509-519

The generalization with respect to domain shifts, as they frequently appear in a pplications such as autonomous driving, is one of the remaining big challenges f or deep learning models. Therefore, we propose an intra-source style augmentatio n (ISSA) method to improve domain generalization in semantic segmentation. Our m ethod is based on a novel masked noise encoder for StyleGAN2 inversion. The mode l learns to faithfully reconstruct the image preserving its semantic layout thro ugh noise prediction. Random masking of the estimated noise enables the style mi xing capability of our model, i.e. it allows to alter the global appearance with out affecting the semantic layout of an image. Using the proposed masked noise e ncoder to randomize style and content combinations in the training set, ISSA eff ectively increases the diversity of training data and reduces spurious correlati on. As a result, we achieve up to 12.4% mIoU improvements on driving-scene seman tic segmentation under different types of data shifts, i.e., changing geographic locations, adverse weather conditions, and day to night. ISSA is model-agnostic and straightforwardly applicable with CNNs and Transformers. It is also complem entary to other domain generalization techniques, e.g., it improves the recent s tate-of-the-art solution RobustNet by 3% mIoU in Cityscapes to Dark Zurich.

**********

RNAS-MER: A Refined Neural Architecture Search With Hybrid Spatiotemporal Operat ions for Micro-Expression Recognition

Monu Verma, Priyanka Lubal, Santosh Kumar Vipparthi, Mohamed Abdel-Mottaleb; Pro ceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (W ACV), 2023, pp. 4770-4779

Existing NAS methods comprise linear connected convolutional operations and used ample search space to search task-driven convolution neural networks (CNN). These CNN models are computationally expensive and diminish the quality of receptive fields for tasks like micro-expression recognition (MER) with limited training samples. Therefore, we proposed a refined neural architecture search strategy to search a tiny CNN architecture for MER. In addition, we introduced a refined hybrid module (RHM) for innerlevel search space and an optimal path explore network (OPEN) for outer-level search. The RHM focuses on discovering optimal cell structures by incorporating a multilateral hybrid spatiotemporal operation space. Also, spatiotemporal attention blocks are embedded to refine the aggregated cell features. The OPEN search space aims to trace an optimal path between the cells to generate a tiny spatiotemporal CNN architecture instead of covering all possible tracks. The aggregate mix of RHM and OPEN search space availed the NAS method to robustly search and design an effective and efficient framework for MER. Compared with contemporary works, experiments reveal that the RNAS-MER is capable of bridging the gap between NAS algorithms and MER tasks. Further, RNAS-MER achieves new state-of-the-art performances on challenging MER benchmarks, including % on CASME-2, % SMIC, % SAMM, and % on COMPOSITE.

**********

HiFormer: Hierarchical Multi-Scale Representations Using Transformers for Medical Image Segmentation

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, Dorit Merhof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6202-6212

Convolutional neural networks (CNNs) have been the consensus for medical image segmentation tasks. However, they inevitably suffer from the limitation in modeling long-range dependencies and spatial correlations due to the nature of convolution operation. Although Transformers were first developed to address this issue, they fail to capture low-level features. In contrast, it is demonstrated that both local and global features are crucial for dense prediction, such as segmenting in challenging contexts. In this paper, we propose HiFormer, a novel method that efficiently bridges a Convolutional neural network and a Transformer for medical image segmentation. Specifically, we design two multi-scale feature representations using the seminal Swin-Transformer module and a CNN-based encoder. To secure a fine fusion of global and local features obtained from the two aforementioned representations, we propose a Double-Level Fusion (DLF) module in the skip connection of the encoder-decoder outline. Extensive experiments on various medical image segmentation datasets demonstrate the effectiveness of HiFormer over other CNN-based, Transformer-based, and hybrid methods in terms of computational complexity, quantitative and qualitative results

**********

Expert-Defined Keywords Improve Interpretability of Retinal Image Captioning

Ting-Wei Wu, Jia-Hong Huang, Joseph Lin, Marcel Worring; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1859-1868

Automatic machine learning-based (ML-based) medical report generation systems for retinal images suffer from a relative lack of interpretability. Hence, such ML-based systems are still not widely accepted. The main reason is that trust is one of the important motivating aspects of interpretability and humans do not trust blindly. Precise technical definitions of interpretability still lack consensus. Hence, it is difficult to make a human-comprehensible ML-based medical report generation system. Heat maps/saliency maps, i.e., post-hoc explanation approaches, are widely used to improve the interpretability of ML-based medical systems. However, they are well known to be problematic. From an ML-based medical model's perspective, the highlighted areas of an image are considered important for making a prediction. However, from a doctor's perspective, even the hottest regions of a heat map contain both useful and non-useful information. Simply localizing the region, therefore, does not reveal exactly what it was in that area that the model considered useful. Hence, the post-hoc explanation-based method relies on humans who probably have a biased nature to decide what a given heat map mig

ht mean. Interpretability boosters, in particular expert-defined keywords, are e ffective carriers of expert domain knowledge and they are human-comprehensible. In this work, we propose to exploit such keywords and a specialized attention-ba sed strategy to build a more human-comprehensible medical report generation syst em for retinal images. Both keywords and the proposed strategy effectively impro ve the interpretability. The proposed method achieves state-of-the-art performan ce under commonly used text evaluation metrics BLEU, ROUGE, CIDEr, and METEOR. P roject website: https://github.com/Jhhuangkay/Expert-defined-Keywords-Improve-In terpretability-of-Retinal-Image-Captioning.

**********

MORGAN: Meta-Learning-Based Few-Shot Open-Set Recognition via Generative Adversa rial Network

Debabrata Pal, Shirsha Bose, Biplab Banerjee, Yogananda Jeppu; Proceedings of th e IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp . 6295-6304

In few-shot open-set recognition (FSOSR) for hyperspectral images (HSI), one maj or challenge arises due to the simultaneous presence of spectrally fine-grained known classes and outliers. Prior research on generative FSOSR cannot handle suc h a situation due to their inability to approximate the open space prudently. To address this issue, we propose a method, Meta-learning-based Open-set Recogniti on via Generative Adversarial Network (MORGAN), that can learn a finer separatio n between the closed and the open spaces. MORGAN seeks to generate class-conditi oned adversarial samples for both the closed and open spaces in the few-shot reg ime using two GANs by judiciously tuning noise variance while ensuring discrimin ability using a novel Anti-Overlap Latent (AOL) regularizer. Adversarial samples from low noise variance amplify known class data density, and we use samples fr om high noise variance to augment known-unknowns. A first-order episodic strateg y is adapted to ensure stability in the GAN training. Finally, we introduce a co mbination of metric losses which push these augmented known-unknowns or outliers to disperse in the open space while condensing known class distributions. Exten sive experiments on four benchmark HSI datasets indicate that MORGAN achieves st ate-of-the-art FSOSR performance consistently.

**********

Fine-Grained Affordance Annotation for Egocentric Hand-Object Interaction Videos

Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, Yoichi Sato ; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi on (WACV), 2023, pp. 2155-2163

Object affordance is an important concept in hand-object interaction, providing information on action possibilities based on human motor capacity and objects' p hysical property thus benefiting tasks such as action anticipation and robot imi tation learning. However, the definition of affordance in existing datasets ofte n: 1) mix up affordance with object functionality; 2) confuse affordance with go al-related action; and 3) ignore human motor capacity. This paper proposes an ef ficient annotation scheme to address these issues by combining goal-irrelevant m otor actions and grasp types as affordance labels and introducing the concept of mechanical action to represent the action possibilities between two objects. We provide new annotations by applying this scheme to the EPIC-KITCHENS dataset an d test our annotation with tasks such as affordance recognition, hand-object int eraction hotspots prediction, and cross-domain evaluation of affordance. The res ults show that models trained with our annotation can distinguish affordance fro m other concepts, predict fine-grained interaction possibilities on objects, and generalize through different domains.

**********

End-to-End Single-Frame Image Signal Processing for High Dynamic Range Scenes

Khanh Quoc Dinh, Kwang Pyo Choi; Proceedings of the IEEE/CVF Winter Conference o n Applications of Computer Vision (WACV), 2023, pp. 2449-2458

This paper considers photography of high dynamic range scenes containing mixture s of shadows and highlights on mobile phones. Multi-frame merging constructs a h igh-quality image at the cost of capturing multiple frames of the same scene. Co ntrarily, end-to-end optimized image signal processing (E2EISP) produces an enha

nced image from a single-frame Bayer array. This paper combines the merits of th e two approaches by using labels of high-quality multi-frame merged images to tr ain E2EISP with a novel neural network architecture composed of a multi-head mix ture of brightness enhancement for accurately processing shadows/highlights and a multi-head mixture of image processing featured camera settings of white balan ce and color correction for a proper color generation. We also proposed a combin ation of supervised, unsupervised, and generative adversarial losses for brightn ess, edge, and detail enhancement. Experimental results show that the proposed s ingle-frame ISP produces enhanced images and outperforms state-of-the-art method s.
**********

Indirect Adversarial Losses via an Intermediate Distribution for Training GANs
Rui Yang, Duc Minh Vo, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Confe rence on Applications of Computer Vision (WACV), 2023, pp. 4652-4661
In this study, we consider the weak convergence characteristics of the Integral Probability Metrics (IPM) methods in training Generative Adversarial Networks (G ANs). We first concentrate on a successful IPM-based GAN method that employs a r epulsive version of the Maximum Mean Discrepancy (MMD) as the discriminator loss (called repulsive MMD-GAN). We reinterpret its repulsive metrics as an indirect discriminator loss function toward an intermediate distribution. This allows us to propose a novel generator loss via such an intermediate distribution based o n our reinterpretation. Our indirect adversarial losses use a simple known distr ibution (i.e., the Normal or Uniform distribution in our experiments) to simulat e indirect adversarial learning between three parts -- real, fake, and intermedi ate distributions. Furthermore, we found the Kernelized Stein Discrepancy (KSD) from the IPM family as the adversarial loss function to avoid randomness from in termediate distribution samples because the target side (intermediate one) is sa mple-free in KSD. Experiments on several real-world datasets show that our metho ds can successfully train GANs with the intermediate-distribution-based KSD and MMD and can outperform previous loss metrics.
**********

GAF-Net: Improving the Performance of Remote Sensing Image Fusion Using Novel Gl obal Self and Cross Attention Learning
Ankit Jha, Shirsha Bose, Biplab Banerjee; Proceedings of the IEEE/CVF Winter Con ference on Applications of Computer Vision (WACV), 2023, pp. 6354-6363
The notion of self and cross-attention learning has been found to substantially boost the performance of remote sensing (RS) image fusion. However, while the se lf-attention models fail to incorporate the global context due to the limited si ze of the receptive fields, cross-attention learning may generate ambiguous feat ures as the feature extractors for all the modalities are jointly trained. This results in the generation of redundant multi-modal features, thus limiting the f usion performance. To address these issues, we propose a novel fusion architectu re called Global Attention based Fusion Network (GAF-Net), equipped with novel s elf and cross-attention learning techniques. We introduce the within-modality fe ature refinement module through global spectral-spatial attention learning using the query-key-value processing where both the global spatial and channel contex ts are used to generate two channel attention masks. Since it is non-trivial to generate the cross-attention from within the fusion network, we propose to lever age two auxiliary tasks of modality-specific classification to produce highly di scriminative cross-attention masks. Finally, to ensure non-redundancy, we propos e to penalize the high correlation between attended modality-specific features. Our extensive experiments on five benchmark datasets, including optical, multisp ectral (MS), hyperspectral (HSI), light detection and ranging (LiDAR), synthetic aperture radar (SAR), and audio modalities establish the superiority of GAF-Net concerning the literature.
**********

From Forks to Forceps: A New Framework for Instance Segmentation of Surgical Ins truments
Britty Baby, Daksh Thapar, Mustafa Chasmai, Tamajit Banerjee, Kunal Dargan, Ashi sh Suri, Subhashis Banerjee, Chetan Arora; Proceedings of the IEEE/CVF Winter Co

nference on Applications of Computer Vision (WACV), 2023, pp. 6191-6201
Minimally invasive surgeries and related applications demand surgical tool class ification and segmentation at the instance level. Surgical tools are similar in appearance and are long, thin, and handled at an angle. The fine-tuning of state -of-the-art (SOTA) instance segmentation models trained on natural images for in strument segmentation has difficulty discriminating instrument classes. Our rese arch demonstrates that while the bounding box and segmentation mask are often ac curate, the classification head misclassifies the class label of the surgical in strument. We present a new neural network framework that adds a classification m odule as a new stage to existing instance segmentation models. This module speci alizes in improving the classification of instrument masks generated by the exis ting model. The module comprises multi-scale mask attention, which attends to th e instrument region and masks the distracting background features. We propose tr aining the proposed classifier module using metric learning with arc loss to han dle low inter-class variance of surgical instruments. We conduct exhaustive expe riments on the benchmark datasets EndoVis2017 and EndoVis2018. We demonstrate th at our method outperforms all (more than 18) SOTA methods compared with and impr oves the \sota performance by at least 12 points (20%) on the EndoVis2017 benchm ark challenge and generalizes effectively across the datasets. Project page with source code is available at nets-iitd.github.io/s3net.
**********

Harnessing Unrecognizable Faces for Improving Face Recognition
Siqi Deng, Yuanjun Xiong, Meng Wang, Wei Xia, Stefano Soatto; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3424-3433
The common implementation of face recognition systems as a cascade of a detectio n stage and a recognition or verification stage can cause problems beyond failur es of the detector. When the detector succeeds, it can detect faces that cannot be recognized, no matter how capable the recognition system is. Recognizability, a latent variable, should therefore be factored into the design and implementat ion of face recognition systems. We propose a measure of recognizability of a fa ce image that leverages a key empirical observation: An embedding of face images , implemented by a deep neural network trained using mostly recognizable identit ies, induces a partition of the hypersphere whereby unrecognizable identities cl uster together. This occurs regardless of the phenomenon that causes a face to b e unrecognizable, be it optical or motion blur, partial occlusion, spatial quant ization, or poor illumination. Therefore, we use the distance from such an "unre cognizable identity" as a measure of recognizability, and incorporate it into th e design of the overall system. We show that accounting for recognizability redu ces the error rate of single-image face recognition by 58% at FAR=1e-5 on the IJ B-C Covariate Verification benchmark, and reduces the verification error rate by 24% at FAR=1e-5 in set-based recognition on the IJB-C benchmark.
**********

Multi-View Action Recognition Using Contrastive Learning
Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M. de Melo, Rama Chellappa; Procee dings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV ), 2023, pp. 3381-3391
In this work, we present a method for RGB-based action recognition using multi-v iew videos. We present a supervised contrastive learning framework to learn a fe ature embedding robust to changes in viewpoint, by effectively leveraging multi- view data. We use an improved supervised contrastive loss and augment the positi ves with those coming from synchronized viewpoints. We also propose a new approa ch to use classifier probabilities to guide the selection of hard negatives in t he contrastive loss, to learn a more discriminative representation. Negative sam ples from confusing classes based on posterior are weighted higher. We also show that our method leads to better domain generalization compared to the standard supervised training based on synthetic multi-view data. Extensive experiments on real (NTU-60, NTU-120, NUMA) and synthetic (RoCoG) data demonstrate the effecti veness of our approach.
**********

TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation
Jinyu Yang, Jingjing Liu, Ning Xu, Junzhou Huang; Proceedings of the IEEE/CVF Wi
nter Conference on Applications of Computer Vision (WACV), 2023, pp. 520-530
Unsupervised domain adaptation (UDA) aims to transfer the knowledge learnt from
a labeled source domain to an unlabeled target domain. Previous work is mainly b
uilt upon convolutional neural networks (CNNs) to learn domain-invariant represe
ntations. With the recent exponential increase in applying Vision Transformer (V
iT) to vision tasks, the capability of ViT in adapting cross-domain knowledge, h
owever, remains unexplored in the literature. To fill this gap, this paper first
 comprehensively investigates the performance of ViT on a variety of domain adap
tation tasks. Surprisingly, ViT demonstrates superior generalization ability, wh
ile the performance can be further improved by incorporating adversarial adaptat
ion. Notwithstanding, directly using CNNs-based adaptation strategies fails to t
ake the advantage of ViT's intrinsic merits (e.g., attention mechanism and seque
ntial image representation) which play an important role in knowledge transfer.
To remedy this, we propose an unified framework, namely Transferable Vision Tran
sformer (TVT), to fully exploit the transferability of ViT for domain adaptation
. Specifically, we delicately devise a novel and effective unit, which we term T
ransferability Adaption Module (TAM). By injecting learned transferabilities int
o attention blocks, TAM compels ViT focus on both transferable and discriminativ
e features. Besides, we leverage discriminative clustering to enhance feature di
versity and separation which are undermined during adversarial domain alignment.
 To verify its versatility, we perform extensive studies of TVT on four benchmar
ks and the experimental results demonstrate that TVT attains significant improve
ments compared to existing state-of-the-art UDA methods.
**********
Interpolated SelectionConv for Spherical Images and Surfaces
David Hart, Michael Whitney, Bryan Morse; Proceedings of the IEEE/CVF Winter Con
ference on Applications of Computer Vision (WACV), 2023, pp. 321-330
We present a new and general framework for convolutional neural network operatio
ns on spherical (or omnidirectional) images. Our approach represents the surface
 as a graph of connected points that doesn't rely on a particular sampling strat
egy. Additionally, by using an interpolated version of SelectionConv, we can ope
rate on the sphere while using existing 2D CNNs and their weights. Since our met
hod leverages existing graph implementations, it is also fast and can be fine-tu
ned efficiently. Our method is also general enough to be applied to any surface
type, even those that are topologically non-simple. We demonstrate the effective
ness of our technique on the tasks of style transfer and segmentation for sphere
s as well as stylization for 3D meshes. We provide a thorough ablation study of
the performance of various spherical sampling strategies.
**********
Match Cutting: Finding Cuts With Smooth Visual Transitions
Boris Chen, Amir Ziai, Rebecca S. Tucker, Yuchen Xie; Proceedings of the IEEE/CV
F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2115-21
25
A match cut is a transition between a pair of shots that uses similar framing, c
omposition, or action to fluidly bring the viewer from one scene to the next. Ma
tch cuts are frequently used in film, television, and advertising. However, find
ing shots that work together is a highly manual and time-consuming process that
can take days. We propose a modular and flexible system to efficiently find high
-quality match cut candidates starting from millions of shot pairs. We annotate
and release a dataset of approximately 20,000 labeled pairs that we use to evalu
ate our system, using both classification and metric learning approaches that le
verage a variety of image, video, audio, and audio-visual feature extractors. In
 addition, we release code and embeddings for reproducing our experiments at git
hub.com/netflix/matchcut.
**********
Are Straight-Through Gradients and Soft-Thresholding All You Need for Sparse Tra
ining?
Antoine Vanderschueren, Christophe De Vleeschouwer; Proceedings of the IEEE/CVF

Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3808-3817
Turning the weights to zero when training a neural network helps in reducing the computational complexity at inference. To progressively increase the sparsity ratio in the network without causing sharp weight discontinuities during training, our work combines soft-thresholding and straight-through gradient estimation to update the raw, i.e. non-thresholded, version of zeroed weights. Our method, named ST-3 for straight-through/soft-thresholding/sparse-training, obtains SoA results, both in terms of accuracy/sparsity and accuracy/FLOPS trade-offs, when progressively increasing the sparsity ratio in a single training cycle. In particular, despite its simplicity, ST-3 favorably compares to the most recent methods, adopting differentiable formulations or bio-inspired neuroregeneration principles. This suggests that the key ingredients for effective sparsification primarily lie in the ability to give the weights the freedom to evolve smoothly across the zero state while progressively increasing the sparsity ratio. Source code and weights available at https://github.com/vanderschuea/stthree.
**********

Fast and Accurate: Video Enhancement Using Sparse Depth
Yu Feng, Patrick Hansen, Paul N. Whatmough, Guoyu Lu, Yuhao Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4492-4500
This paper presents a general framework to build fast and accurate algorithms for video enhancement tasks such as super-resolution, deblurring, and denoising. Essential to our framework is the realization that the accuracy, rather than the density, of pixel flows is what is required for high-quality video enhancement. Most of prior works take the opposite approach: they estimate dense (per-pixel)--but generally less robust--flows, mostly using computationally costly algorithms. Instead, we propose a lightweight flow estimation algorithm; it fuses the sparse point cloud data and (even sparser and less reliable) IMU data available in modern autonomous agents to estimate the flow information. Building on top of the flow estimation, we demonstrate a general framework that integrates the flows in a plug-and-play fashion with different task-specific layers. Algorithms built in our framework achieve 1.78x -- 187.41x speedup while providing a 0.42dB - 6.70 dB quality improvement over competing methods.
**********

SGPCR: Spherical Gaussian Point Cloud Representation and Its Application To Object Registration and Retrieval
Driton Salihu, Eckehard Steinbach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 572-581
Retrieving and aligning CAD models from databases with scanned real-world point clouds remains an important topic for 3D reconstruction. Due to zero point-to-point correspondences between the sampled CAD model and the scanned real-world object, an information-rich representation of point clouds is needed. We propose SGPCR, a novel method for representing 3D point clouds by Spherical Gaussians for efficient, stable, and rotation-equivariant representation. We also propose a rotation-invariant convolution to improve the representation quality through a trainable optimization process. In addition, we demonstrate the strengths of SGPCR-based point cloud representation using the fundamental challenge of shape retrieval and point cloud registration on point clouds with zero point-to-point correspondences. Under these conditions, our approach improves registration quality by reducing chamfer distance by up to 90% and rotation root mean square error by up to 86% compared to the state of the art. Furthermore, the proposed SGCPR is used for one-shot shape retrieval and registration and improves retrieval precision by up to 58% over comparable methods.
**********

Nested Deformable Multi-Head Attention for Facial Image Inpainting
Shruti S. Phutke, Subrahmanyam Murala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6078-6087
Extracting adequate contextual information is an important aspect of any image inpainting method. To achieve this, ample image inpainting methods are available that aim to focus on large receptive fields. Recent advancements in the deep lea

rning field with the introduction of transformers for image inpainting paved the way toward plausible results. Stacking multiple transformer blocks in a single layer causes the architecture to become computationally complex. In this context, we propose a novel lightweight architecture with a nested deformable attention based transformer layer for feature fusion. The nested attention helps the network to focus on long-term dependencies from encoder and decoder features. Also, multi head attention consisting of a deformable convolution is proposed to delve into the diverse receptive fields. With the advantage of nested and deformable attention, we propose a lightweight architecture for facial image inpainting. The results comparison on Celeb HQ [25] dataset using known (NVIDIA) and unknown (QD-IMD) masks and Places2 [57] dataset with NVIDIA masks along with extensive ablation study prove the superiority of the proposed approach for image inpainting tasks. The code is available at: https://github.com/shrutiphutke/NDMA_ Facial_Inpainting.

**********

Fashion Image Retrieval With Text Feedback by Additive Attention Compositional Learning

Yuxin Tian, Shawn Newsam, Kofi Boakye; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1011-1021

Effective fashion image retrieval with text feedback stands to impact a range of real-world applications, such as e-commerce. Given a source image and text feedback that describes the desired modifications to that image, the goal is to retrieve the target images that resemble the source yet satisfy the given modifications by composing a multi-modal (image-text) query. We propose a novel solution to this problem, Additive Attention Compositional Learning (AACL), that uses a multi-modal transformer-based architecture and effectively models the image-text contexts. Specifically, we propose a novel image-text composition module based on additive attention that can be seamlessly plugged into deep neural networks. We also introduce a new challenging benchmark derived from the Shopping100k dataset. AACL is evaluated on three large-scale datasets (FashionIQ, Fashion200k, and Shopping100k), each with strong baselines. Extensive experiments show that AACL achieves new state-of-the-art results on all three datasets.

**********

Center-Aware Adversarial Augmentation for Single Domain Generalization

Tianle Chen, Mahsa Baktashmotlagh, Zijian Wang, Mathieu Salzmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4157-4165

Domain generalization (DG) aims to learn a model from multiple training (i.e., source) domains that can generalize well to the unseen test (i.e., target) data coming from a different distribution. Single domain generalization (Single-DG) has recently emerged to tackle a more challenging, yet realistic setting, where only one source domain is available at training time. The existing Single-DG approaches typically are based on data augmentation strategies and aim to expand the span of source data by augmenting out-of-domain samples. Generally speaking, they aim to generate hard examples to confuse the classifier. While this may make the classifier robust to small perturbation, the generated samples are typically not diverse enough to mimic a large domain shift, resulting in sub-optimal generalization performance To alleviate this, we propose a center-aware adversarial augmentation technique that expands the source distribution by altering the source samples so as to push them away from the class centers via a novel angular center loss. We conduct extensive experiments to demonstrate the effectiveness of our approach on several benchmark datasets for Single-DG and show that our method outperforms the state-of-the-art in most cases.

**********

SD-Pose: Structural Discrepancy Aware Category-Level 6D Object Pose Estimation

Guowei Li, Dongchen Zhu, Guanghui Zhang, Wenjun Shi, Tianyu Zhang, Xiaolin Zhang, Jiamao Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5685-5694

Category-level 6D object pose estimation aims to predict the full pose and size information for previously unseen instances from known categories, which is an e

ssential portion of robot grasping and augmented reality. However, the core chal lenge of this task still is the enormous shape variation within each category. W ith regard to the challenge, we propose a novel framework SD-Pose, which utilize s the instance-category structural discrepancy and the potential geometric-seman tic association to enhance the exploration of the intra-class shape information. Specifically, an information exchange augmentation (IEA) module is introduced t o supplement the instance-category structural information by their structural di screpancy, thus facilitating the enhanced geometric information to contain both the character of instance shape and the commonality of category structure. For c omplementing the deficiencies of structural information adaptively, a semantic d ynamic fusion (SDF) module is further designed to fuse semantic and geometric fe atures. Finally, the proposed SD-Pose framework equipped with the IEA and SDF mo dules hierarchically supplements instance-category structural information in a s tacked manner and achieves state-of-the-art performance on the CAMERA25 and REAL 275 datasets.

**********

FLOAT: Fast Learnable Once-for-All Adversarial Training for Tunable Trade-Off Be tween Accuracy and Robustness

Souvik Kundu, Sairam Sundaresan, Massoud Pedram, Peter A. Beerel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2349-2358

Existing models that achieve state-of-the-art (SOTA) performance on both clean a nd adversarially-perturbed images rely on convolution operations conditioned wit h feature-wise linear modulation (FiLM) layers. These layers require additional parameters and are hyperparameter sensitive. They significantly increase trainin g time, memory cost, and potential latency which can be costly for resource-limi ted or real-time applications. In this paper, we present a fast learnable once-f or-all adversarial training (FLOAT) algorithm, which instead of the existing FiL M-based conditioning, presents a unique weight conditioned learning that require s no additional layer, thereby incurring no significant increase in parameter co unt, training time, or network latency compared to standard adversarial training . In particular, we add configurable scaled noise to the weight tensors that ena bles a trade-off between clean and adversarial performance. Extensive experiment s show that FLOAT can yield SOTA performance improving both clean and perturbed image classification by up to 6% and 10%, respectively. Moreover, real hardwar e measurement shows that FLOAT can reduce the training time by up to 1.43x with fewer model parameters of up to 1.47x on iso-hyperparameter settings compared to the FiLM-based alternatives. Additionally, to further improve memory efficiency we introduce FLOAT sparse (FLOATS), a form of non-iterative model pruning, and provide detailed empirical analysis in yielding a three-way accuracy-robustness-complexity trade-off for these new class of pruned conditionally trained models.

**********

Spatio-Temporal Action Detection Under Large Motion

Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu, Luc Van Gool; Proceedin gs of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6009-6018

Current methods for spatiotemporal action tube detection often extend a bounding box proposal at a given key-frame into a 3D temporal cuboid and pool features f rom nearby frames. However, such pooling fails to accumulate meaningful spatiote mporal features if the position or shape of the actor shows large 2D motion and variability through the frames, due to large camera motion, large actor shape de formation, fast actor action and so on. In this work, we aim to study the perfor mance of cuboid-aware feature aggregation in action detection under large action . Further, we propose to enhance actor feature representation under large motion by tracking actors and performing temporal feature aggregation along the respec tive tracks. We define the actor motion with intersection-over-union (IoU) betwe en the boxes of action tubes/tracks at various fixed time scales. The action hav ing a large motion would result in lower IoU over time, and slower actions would maintain higher IoU. We find that track-aware feature aggregation consistently achieves a large improvement in action detection performance, especially for act

ions under large motion compared to the cuboid-aware baseline. As a result, we a lso report state-of-the-art on the large-scale MultiSports dataset.
**********

SLI-pSp: Injecting Multi-Scale Spatial Layout in pSp
Aradhya Neeraj Mathur, Anish Madan, Ojaswa Sharma; Proceedings of the IEEE/CVF W inter Conference on Applications of Computer Vision (WACV), 2023, pp. 4095-4104
We propose SLI-pSp, a general purpose Image-to-Image (I2I) translation model tha t encodes spatial layout information as well as style in the generator, using pS p as the base architecture. Previous methods like pSp have shown promising resul ts by leveraging StyleGAN as a generator in various I2I tasks but they seem to m iss finer or under-represented details in facial images like earrings and caps, and break down on complex datasets due to their solely global approach. To addre ss these shortcomings, we propose a technique termed Spatial Layout Injection (S LI-pSp) that encodes spatial layout information in the input image in the StyleG AN generator along with style. We do so without modifying the style vector injec tion in the generator through pSp's map2style network, but rather by combining S LI with noise layers in the StyleGAN generator at multiple spatial scales. Such an approach helps preserve global aspects of image generation as well as enhance spatial layout details in the output. We experiment on several challenging data sets and across several I2I tasks that highlight the effectiveness of our approa ch over previous methods with respect to finer details in the generated image an d overall visual quality.
**********

Anomaly Clustering: Grouping Images Into Coherent Clusters of Anomaly Types
Kihyuk Sohn, Jinsung Yoon, Chun-Liang Li, Chen-Yu Lee, Tomas Pfister; Proceeding s of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2 023, pp. 5479-5490
We study anomaly clustering, grouping data into coherent clusters of anomaly typ es. This is different from anomaly detection that aims to divide anomalies from normal data. Unlike object-centered image clustering, anomaly clustering is part icularly challenging as anomalous patterns are subtle and local. We present a si mple yet effective clustering framework using a patch-based pretrained deep embe ddings and off-the-shelf clustering methods. We define a distance function betwe en images, each of which is represented as a bag of embeddings, by the Euclidean distance between weighted averaged embeddings. The weight defines the importanc e of instances (i.e., patch embeddings) in the bag, which may highlight defectiv e regions. We compute weights in an unsupervised way or in a semi-supervised way when labeled normal data is available. Extensive experimental studies show the effectiveness of the proposed clustering framework along with a novel distance f unction upon exist-ing multiple instance or deep clustering frameworks. Over-all , our framework achieves 0.451 and 0.674 normalized mutual information scores on MVTec object and texture categories and further improve with a few labeled norm al data (0.577, 0.669), far exceeding the baselines (0.244, 0.273) or state-of-t he-art deep clustering methods (0.176, 0.277).
**********

DSAG: A Scalable Deep Framework for Action-Conditioned Multi-Actor Full Body Mot ion Synthesis
Debtanu Gupta, Shubh Maheshwari, Sai Shashank Kalakonda, Manasvi Vaidyula, Ravi Kiran Sarvadevabhatla; Proceedings of the IEEE/CVF Winter Conference on Applicat ions of Computer Vision (WACV), 2023, pp. 4300-4308
We introduce DSAG, a controllable deep neural framework for action-conditioned g eneration of full body multi-actor variable duration actions. To compensate for incompletely detailed finger joints in existing large-scale datasets, we introduc e full body dataset variants with detailed finger joints. To overcome shortcomi ngs in existing generative approaches, we introduce dedicated representations fo r encoding finger joints. We also introduce novel spatiotemporal transformation blocks with multi-head self-attention and specialized temporal processing. The d esign choices enable generations for a large range in body joint counts (24 - 52 ), frame rates (13 - 50), global body movement (in-place, locomotion) and action categories (12 - 120), across multiple datasets (NTU-120, HumanAct12, UESTC, Hu

man3.6M). Our experimental results demonstrate DSAG's significant improvements o
ver state-of-the-art, its suitability for action-conditioned generation at scale
.
**********

RIFT: Disentangled Unsupervised Image Translation via Restricted Information Flo
w

Ben Usman, Dina Bashkirova, Kate Saenko; Proceedings of the IEEE/CVF Winter Conf
erence on Applications of Computer Vision (WACV), 2023, pp. 2420-2429

Unsupervised image-to-image translation methods aim to map images from one domai
n into plausible examples from another domain while preserving the structure sha
red across two domains. In the many-to-many setting, an additional guidance exam
ple from the target domain is used to determine the domain-specific factors of v
ariation of the generated image. In the absence of attribute annotations, method
s have to infer which factors of variation are specific to each domain from data
 during training. In this paper, we show that many state-of-the-art architecture
s implicitly treat textures and colors as always being domain-specific, and thus
 fail when they are not. We propose a new method called RIFT that does not rely
on such inductive architectural biases and instead infers which attributes are d
omain-specific vs shared directly from data. As a result, RIFT achieves consiste
ntly high cross-domain manipulation accuracy across multiple datasets spanning a
 wide variety of domain-specific and shared factors of variation.
**********

Adversarial Robustness in Discontinuous Spaces via Alternating Sampling & Descen
t

Rahul Venkatesh, Eric Wong, Zico Kolter; Proceedings of the IEEE/CVF Winter Conf
erence on Applications of Computer Vision (WACV), 2023, pp. 4662-4671

Several works have shown that deep learning models are vulnerable to adversarial
 attacks where seemingly simple label-preserving changes to the input image lead
 to incorrect predictions. To combat this, gradient based adversarial training i
s generally employed as a standard defense mechanism. However, in cases where th
e loss landscape is discontinuous with respect to a given perturbation set, firs
t order methods get stuck in local optima, and fail to defend against threat. Th
is is often a problem for many physically realizable perturbation sets such as 2
D affine transformations and 3D scene parameters. To work in such settings, we i
ntroduce a new optimization framework that alternates between global zeroth orde
r sampling and local gradient updates to compute strong adversaries that can be
used to harden the model against attack. Further, we design a powerful optimizat
ion algorithm using this framework, called Alternating Evolutionary Sampling and
 Descent (ASD), which combines an evolutionary search strategy (viz. covariance
matrix adaptation) with gradient descent. We consider two settings with disconti
nuous/discrete and non-convex loss landscapes to evaluate ASD: a) 3D scene param
eters and b) 2D patch attacks, and find that it achieves state-of-the-art result
s on adversarial robustness.
**********

Backprop Induced Feature Weighting for Adversarial Domain Adaptation With Iterat
ive Label Distribution Alignment

Thomas Westfechtel, Hao-Wei Yeh, Qier Meng, Yusuke Mukuta, Tatsuya Harada; Proce
edings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC
V), 2023, pp. 392-401

The requirement for large labeled datasets is one of the limiting factors for tr
aining accurate deep neural networks. Unsupervised domain adaptation tackles thi
s problem of limited training data by transferring knowledge from one domain, wh
ich has many labeled data, to a different domain for which little to no labeled
data is available. One common approach is to learn domain-invariant features for
 example with an adversarial approach. Previous methods often train the domain c
lassifier and label classifier network separately, where both classification net
works have little interaction with each other. In this paper, we introduce a cla
ssifier-based backprop-induced weighting of the feature space. This approach has
 two main advantages. Firstly, it lets the domain classifier focus on features t
hat are important for the classification, and, secondly, it couples the classifi

cation and adversarial branch more closely. Furthermore, we introduce an iterati ve label distribution alignment method, that employs results of previous runs to approximate a class-balanced dataloader. We conduct experiments and ablation st udies on three benchmarks Office-31, OfficeHome, and DomainNet to show the effec tiveness of our proposed algorithm.

**********

Understanding the Role of Mixup in Knowledge Distillation: An Empirical Study
Hongjun Choi, Eun Som Jeon, Ankita Shukla, Pavan Turaga; Proceedings of the IEEE /CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2319 -2328

Mixup is a popular data augmentation technique based on creating new samples by linear interpolation between two given data samples, to improve both the general ization and robustness of the trained model. Knowledge distillation (KD), on the other hand, is widely used for model compression and transfer learning, which i nvolves using a larger network's implicit knowledge to guide the learning of a s maller network. At first glance, these two techniques seem very different, howev er, we found that "smoothness" is the connecting link between the two and is als o a crucial attribute in understanding KD's interplay with mixup. Although many mixup variants and distillation methods have been proposed, much remains to be u nderstood regarding the role of a mixup in knowledge distillation. In this paper , we present a detailed empirical study on various important dimensions of compa tibility between mixup and knowledge distillation. We also scrutinize the behavi or of the networks trained with a mixup in the light of knowledge distillation t hrough extensive analysis, visualizations, and comprehensive experiments on imag e classification. Finally, based on our findings, we suggest improved strategies to guide the student network to enhance its effectiveness. Additionally, the fi ndings of this study provide insightful suggestions to researchers and practitio ners that commonly use techniques from KD. Our code is available at https://gith ub.com/hchoi71/MIX-KD.

**********

Revisiting Training-Free NAS Metrics: An Efficient Training-Based Method
Taojiannan Yang, Linjie Yang, Xiaojie Jin, Chen Chen; Proceedings of the IEEE/CV F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4751-47 60

Recent neural architecture search (NAS) works proposed training-free metrics to rank networks which largely reduced the search cost in NAS. In this paper, we re visit these training-free metrics and find that: (1) the number of parameters (# Param), which is the most straightforward training-free metric, is overlooked in previous works but is surprisingly effective, (2) recent training-free metrics largely rely on the #Param information to rank networks. Our experiments show th at the performance of recent training-free metrics drops dramatically when the # Param information is not available. Motivated by these observations, we argue th at metrics less correlated with the #Param are desired to provide additional inf ormation for NAS. We propose a light-weight training-based metric which has a we ak correlation with the #Param while achieving better performance than training- free metrics at a lower search cost. Specifically, on DARTS search space, our me thod completes searching directly on ImageNet in only 2.6 GPU hours and achieves a top-1/top-5 error rate of 24.1%/7.1%, which is competitive among state-of-the -art NAS methods.

**********

MFCFlow: A Motion Feature Compensated Multi-Frame Recurrent Network for Optical Flow Estimation
Yonghu Chen, Dongchen Zhu, Wenjun Shi, Guanghui Zhang, Tianyu Zhang, Xiaolin Zha ng, Jiamao Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5068-5077

Occlusions have long been a hard nut to crack in optical flow estimation due to ambiguous pixels matching between abutting images. Current methods only take two consecutive images as input, which is challenging to capture temporal coherence and reason about occluded regions. In this paper, we propose a novel optical fl ow estimation framework, namely MFCFlow, which attempts to compensate for the in

formation of occlusions by mining and transferring motion features between multiple frames. Specifically, we construct a Motion-guided Feature Compensation cell (MFC cell) to enhance the ambiguous motion features according to the correlation of previous features obtained by attention-based structure. Furthermore, a Top K attention strategy is developed and embedded into the MFC cell to improve the subsequent matching quality. Extensive experiments demonstrate that our MFCFlow achieves significant improvements in occluded regions and attains state-of-the-art performances on both Sintel and KITTI benchmarks among other multi-frame optical flow methods.
**********

Mapping DNN Embedding Manifolds for Network Generalization Prediction
Molly O'Brien, Brett Wolfinger, Julia Bukowski, Mathias Unberath, Aria Pezeshk, Gregory D. Hager; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6524-6533
Deep Neural Networks(DNN) often fail in surprising ways, and predicting how well a trained DNN will generalize in a new, external operating domain is essential for deploying DNNs in safety critical applications, e.g., perception for self-driving vehicles or medical image analysis. Recently, the task of Network Generalization Prediction (NGP) has been proposed to predict how a DNN will generalize in an external operating domain. Previous NGP approaches have leveraged multiple labeled test sets or labeled metadata. In this study, we propose an embedding map, the first NGP approach that predicts DNN performance based on how unlabeled images from an external operating domain map in the DNN embedding space. We evaluate our proposed Embedding Map and other recently proposed NGP approaches for pedestrian, melanoma, and animal classification tasks. We find that our embedding map has the best average NGP performance, and that our embedding map is effective at modeling complex, non-linear embedding space structures.
**********

GliTr: Glimpse Transformers With Spatiotemporal Consistency for Online Action Prediction
Samrudhdhi B. Rangrej, Kevin J. Liang, Tal Hassner, James J. Clark; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3413-3423
Many online action prediction models observe complete frames to locate and attend to informative subregions in the frames called glimpses and recognize an ongoing action based on global and local information. However, in applications with constrained resources, an agent may not be able to observe the complete frame, yet must still locate useful glimpses to predict an incomplete action based on local information only. In this paper, we develop Glimpse Transformers (GliTr), which observe only narrow glimpses at all times, thus predicting an ongoing action and the following most informative glimpse location based on the partial spatiotemporal information collected so far. In the absence of a ground truth for the optimal glimpse locations for action recognition, we train GliTr using a novel spatiotemporal consistency objective: We require GliTr to attend to the glimpses with features similar to the corresponding complete frames (i.e. spatial consistency) and the resultant class logits at time t equivalent to the ones predicted using whole frames up to t (i.e. temporal consistency). Inclusion of our proposed consistency objective yields 10% higher accuracy on the Something-Something-v2 (SSv2) dataset than the baseline cross-entropy objective. Overall, despite observing only 33% of the total area per frame, GliTr achieves 53.02% and 93.91% accuracy on the SSv2 and Jester datasets, respectively.
**********

ImpDet: Exploring Implicit Fields for 3D Object Detection
Xuelin Qian, Li Wang, Yi Zhu, Li Zhang, Yanwei Fu, Xiangyang Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4260-4270
Conventional 3D object detection approaches concentrate on bounding boxes representation learning with several parameters, i.e., localization, dimension, and orientation. Despite its popularity and universality, such a straightforward paradigm is sensitive to slight numerical deviations, especially in localization. By

exploiting the property that point clouds are naturally captured on the surface of objects along with accurate location and intensity information, we introduce a new perspective that views bounding box regression as an implicit function. This leads to our proposed framework, termed Implicit Detection or ImpDet, which leverages implicit field learning for 3D object detection. Our ImpDet assigns specific values to points in different local 3D spaces, thereby high-quality boundaries can be generated by classifying points inside or outside the boundary. To solve the problem of sparsity on the object surface, we further present a simple yet efficient virtual sampling strategy to not only fill the empty region, but also learn rich semantic features to help refine the boundaries. Extensive experimental results on KITTI and Waymo benchmarks demonstrate the effectiveness and robustness of unifying implicit fields into object detection.

**********

Towards Few-Annotation Learning for Object Detection: Are Transformer-Based Models More Efficient?

Quentin Bouniot, Angélique Loesch, Romaric Audigier, Amaury Habrard; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 75-84

For specialized and dense downstream tasks such as object detection, labeling data requires expertise and can be very expensive, making few-shot and semi-supervised models much more attractive alternatives. While in the few-shot setup we observe that transformer-based object detectors perform better than convolution-based two-stage models for a similar amount of parameters, they are not as effective when used with recent approaches in the semi-supervised setting. In this paper, we propose a semi-supervised method tailored for the current state-of-the-art object detector Deformable DETR in the few-annotation learning setup using a student-teacher architecture, which avoids relying on a sensitive post-processing of the pseudo-labels generated by the teacher model. We evaluate our method on the semi-supervised object detection benchmarks COCO and Pascal VOC, and it outperforms previous methods, especially when annotations are scarce. We believe that our contributions open new possibilities to adapt similar object detection methods in this setup as well.

**********

Domain Adaptive Video Semantic Segmentation via Cross-Domain Moving Object Mixing

Kyusik Cho, Suhyeon Lee, Hongje Seong, Euntai Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 489-498

The network trained for domain adaptation is prone to bias toward the easy-to-transfer classes. Since the ground truth label on the target domain is unavailable during training, the bias problem leads to skewed predictions, forgetting to predict hard-to-transfer classes. To address this problem, we propose Cross-domain Moving Object Mixing (CMOM) that cuts several objects, including hard-to-transfer classes, in the source domain video clip and pastes them into the target domain video clip. Unlike image-level domain adaptation, the temporal context should be maintained to mix moving objects in two different videos. Therefore, we design CMOM to mix with consecutive video frames, so that unrealistic movements are not occurring. We additionally propose Feature Alignment with Temporal Context (FATC) to enhance target domain feature discriminability. FATC exploits the robust source domain features, which are trained with ground truth labels, to learn discriminative target domain features in an unsupervised manner by filtering unreliable predictions with temporal consensus. We demonstrate the effectiveness of the proposed approaches through extensive experiments. In particular, our model reaches mIoU of 53.81% on VIPER -> Cityscapes-Seq benchmark and mIoU of 56.31% on SYNTHIA-Seq -> Cityscapes-Seq benchmark, surpassing the state-of-the-art methods by large margins.

**********

Self-Improving Multiplane-To-Layer Images for Novel View Synthesis

Pavel Solovev, Taras Khakhulin, Denis Korzhenkov; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4309-4318

We present a new method for lightweight novel-view synthesis that generalizes to

an arbitrary forward-facing scene. Recent approaches are computationally expens
ive, require per-scene optimization, or produce a memory-expensive representatio
n. We start by representing the scene with a set of fronto-parallel semitranspar
ent planes and afterwards convert them to deformable layers in an end-to-end man
ner. Additionally, we employ a feed-forward refinement procedure that corrects t
he estimated representation by aggregating information from input views. Our met
hod does not require any fine-tuning when a new scene is processed and can handl
e an arbitrary number of views without any restrictions. Experimental results sh
ow that our approach surpasses recent models in terms of both common metrics and
human evaluation, with the noticeable advantage in inference speed and compactn
ess of the inferred layered geometry.
**********

Patch-Level Gaze Distribution Prediction for Gaze Following
Qiaomu Miao, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE/CVF Winter Con
ference on Applications of Computer Vision (WACV), 2023, pp. 880-889
Gaze following aims to predict where a person is looking in a scene, by predicti
ng the target location, or indicating that the target is located outside the ima
ge. Recent works detect the gaze target by training a heatmap regression task wi
th a pixel-wise mean-square error (MSE) loss, while formulating the in/out predi
ction task as a binary classification task. This training formulation puts a str
ict, pixel-level constraint in higher resolution on the single annotation availa
ble in training, and does not consider annotation variance and the correlation b
etween the two subtasks. To address these issues, we introduce the patch distrib
ution prediction (PDP) method. We replace the in/out prediction branch in previo
us models with the PDP branch, by predicting a patch-level gaze distribution tha
t also considers the outside cases. Experiments show that our model regularizes
the MSE loss by predicting better heatmap distributions on images with larger an
notation variances, meanwhile bridging the gap between the target prediction and
in/out prediction subtasks, showing a significant improvement in performance on
both subtasks on public gaze following datasets.
**********

Self-Distillation for Unsupervised 3D Domain Adaptation
Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, Lu
igi Di Stefano; Proceedings of the IEEE/CVF Winter Conference on Applications of
Computer Vision (WACV), 2023, pp. 4166-4177
Point cloud classification is a popular task in 3D vision. However, previous wor
ks, usually assume that point clouds at test time are obtained with the same pro
cedure or sensor as those at training time. Unsupervised Domain Adaptation (UDA)
instead, breaks this assumption and tries to solve the task on an unlabeled tar
get domain, leveraging only on a supervised source domain. For point cloud class
ification, recent UDA methods try to align features across domains via auxiliary
tasks such as point cloud reconstruction, which however do not optimize the dis
criminative power in the target domain in feature space. In contrast, in this wo
rk, we focus on obtaining a discriminative feature space for the target domain e
nforcing consistency between a point cloud and its augmented version. We then pr
opose a novel iterative self-training methodology that exploits Graph Neural Net
works in the UDA context to refine pseudo-labels. We perform extensive experimen
ts and set the new state-of-the art in standard UDA benchmarks for point cloud c
lassification. Finally, we show how our approach can be extended to more complex
tasks such as part segmentation.
**********

Multivariate Probabilistic Monocular 3D Object Detection
Xuepeng Shi, Zhixiang Chen, Tae-Kyun Kim; Proceedings of the IEEE/CVF Winter Con
ference on Applications of Computer Vision (WACV), 2023, pp. 4281-4290
In autonomous driving, monocular 3D object detection is an important but challen
ging task. Towards accurate monocular 3D object detection, some recent methods r
ecover the distance of objects from the physical height and visual height of obj
ects. Such decomposition framework can introduce explicit constraints on the dis
tance prediction, thus improving its accuracy and robustness. However, the inacc
urate physical height and visual height prediction still may exacerbate the inac

curacy of the distance prediction. In this paper, we improve the framework by mu ltivariate probabilistic modeling. We explicitly model the joint probability dis tribution of the physical height and visual height. This is achieved by learning a full covariance matrix of the physical height and visual height during traini ng, with the guide of a multivariate likelihood. Such explicit joint probability distribution modeling not only leads to robust distance prediction when both th e predicted physical height and visual height are inaccurate, but also brings le arned covariance matrices with expected behaviors. The experimental results on t he challenging Waymo Open and KITTI datasets show the effectiveness of our frame work.
**********

Pixel-Wise Prediction Based Visual Odometry via Uncertainty Estimation
Hao-Wei Chen, Ting-Hsuan Liao, Hsuan-Kung Yang, Chun-Yi Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2518-2528
This paper introduces pixel-wise prediction based visual odometry(PWVO), which i s a dense prediction task that evaluates the values of translation and rotation for every pixel in its input observations. PWVO employs uncertainty estimation t o identify the noisy regions in the input observations, and adopts a selection m echanism to integrate pixel-wise predictions based on the estimated uncertainty maps to derive the final translation and rotation. In order to train PWVO in a c omprehensive fashion, we further develop a data generation workflow for generati ng synthetic training data. The experimental results show that PWVO isable to de liver favorable results. In addition, our analyses validate the effectiveness of the designs adopted in PWVO, and demonstrate that the uncertainty mapsestimated by PWVO is capable of capturing the noises in its input observations.
**********

Relation Preserving Triplet Mining for Stabilising the Triplet Loss In re-Identi fication Systems
Adhiraj Ghosh, Kuruparan Shanmugalingam, Wen-Yan Lin; Proceedings of the IEEE/CV F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4840-48 49
Object appearances change dramatically with pose variations. This creates a chal lenge for embedding schemes that seek to map instances with the same object ID t o locations that are as close as possible. This issue becomes significantly heig htened in complex computer vision tasks such as re-identification(reID). In this paper, we suggest that these dramatic appearance changes are indications that a n object ID is composed of multiple natural groups, and it is counterproductive to forcefully map instances from different groups to a common location. This lea ds us to introduce Relation Preserving Triplet Mining (RPTM), a feature matching guided triplet mining scheme, that ensures that triplets will respect the natur al subgroupings within an object ID. We use this triplet mining mechanism to est ablish a pose-aware, well-conditioned triplet loss by implicitly enforcing view consistency. This allows a single network to be trained with fixed parameters ac ross datasets while providing state-of-the-art results. Code is available at htt ps: //github.com/adhirajghosh/RPTM_reid.
**********

Improving Saliency Models' Predictions of the Next Fixation With Humans' Intrins ic Cost of Gaze Shifts
Florian Kadner, Tobias Thomas, David Hoppe, Constantin A. Rothkopf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 202 3, pp. 2104-2114
The human prioritization of image regions can be modeled in a time invariant fas hion with saliency maps or sequentially with scanpath models. However, while bot h types of models have steadily improved on several benchmarks and datasets, the re is still a considerable gap in predicting human gaze. Here, we leverage two r ecent developments to reduce this gap: theoretical analyses establishing a princ ipled framework for predicting the next gaze target and the empirical measuremen t of the human cost for gaze switches independently of image content. We introdu ce an algorithm in the framework of sequential decision making, which converts a

ny static saliency map into a sequence of dynamic history-dependent value maps, which are recomputed after each gaze shift. These maps are based on 1) a salienc y map provided by an arbitrary saliency model, 2) the recently measured human co st function quantifying preferences in magnitude and direction of eye movements, and 3) a sequential exploration bonus, which changes with each subsequent gaze shift. The parameters of the spatial extent and temporal decay of this explorati on bonus are estimated from human gaze data. The relative contributions of these three components were optimized on the MIT1003 dataset for the NSS score and ar e sufficient to significantly outperform predictions of the next gaze target on NSS and AUC scores for five state of the art saliency models on three image data sets.

**********

Guiding Visual Question Answering With Attention Priors

Thao Minh Le, Vuong Le, Sunil Gupta, Svetha Venkatesh, Truyen Tran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 202 3, pp. 4381-4390

The current success of modern visual reasoning systems is arguably attributed to cross-modality attention mechanisms. However, in deliberative reasoning such as in VQA, attention is unconstrained at each step, and thus may serve as a statis tical pooling mechanism rather than a semantic operation intended to select info rmation relevant to inference. This is because at training time, attention is on ly guided by a very sparse signal (i.e. the answer label) at the end of the infe rence chain. This causes the cross-modality attention weights to deviate from th e desired visual-language bindings. To rectify this deviation, we propose to gui de the attention mechanism using explicit linguistic-visual grounding. This grou nding is derived by connecting structured linguistic concepts in the query to th eir referents among the visual objects. Here we learn the grounding from the pai ring of questions and images alone, without the need for answer annotation or ex ternal grounding supervision. This grounding guides the attention mechanism insi de VQA models through a duality of mechanisms: pre-training attention weight cal culation and directly guiding the weights at inference time on a case-by-case ba sis. The resultant algorithm is capable of probing attention-based reasoning mod els, injecting relevant associative knowledge, and regulating the core reasoning process. This scalable enhancement improves the performance of VQA models, fort ifies their robustness to limited access to supervised data, and increases inter pretability.

**********

Self-Pair: Synthesizing Changes From Single Source for Object Change Detection i n Remote Sensing Imagery

Minseok Seo, Hakjin Lee, Yongjin Jeon, Junghoon Seo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6374-638 3

For change detection in remote sensing, constructing a training dataset for deep learning models is quite difficult due to the requirements of bi-temporal super vision. To overcome this issue, single-temporal supervision which treats change labels as the difference of two semantic masks has been proposed. This novel met hod trains a change detector using two spatially unrelated images with correspon ding semantic labels. However, training with unpaired dataset shows not enough p erformance compared with other methods based on bi-temporal supervision. We susp ect this phenomenon caused by ignorance of meaningful information in the actual bi-temporal pairs.In this paper, we emphasize that the change originates from th e source image and show that manipulating the source image as an after-image is crucial to the performance of change detection. Our method achieves state-of-the -art performance in a large gap than existing methods.

**********

Performer: A Novel PPG-to-ECG Reconstruction Transformer for a Digital Biomarker of Cardiovascular Disease Detection

Ella Lan; Proceedings of the IEEE/CVF Winter Conference on Applications of Compu ter Vision (WACV), 2023, pp. 1991-1999

Electrocardiography (ECG), an electrical measurement which captures cardiac acti

vities, is the gold standard for diagnosing cardiovascular disease (CVD). Howeve r, ECG is infeasible for continuous cardiac monitoring due to its requirement fo r user participation. By contrast, photoplethysmography (PPG) provides easy-to-c ollect data, but its limited accuracy constrains its clinical usage. To combine the advantages of both signals, recent studies incorporate various deep learning techniques for the reconstruction of PPG signals to ECG; however, the lack of c ontextual information as well as the limited abilities to denoise biomedical sig nals ultimately constrain model performance. In this research, we propose Perfor mer, a novel Transformer-based architecture that reconstructs ECG from PPG and c ombines the PPG and reconstructed ECG as multiple modalities for CVD detection. This method is the first time that Transformer sequence-to-sequence translation has been performed on biomedical waveform reconstruction, combining the advantag es of both PPG and ECG. We also create Shifted Patch-based Attention (SPA), an e ffective method to encode/decode the biomedical waveforms. Through fetching the various sequence lengths and capturing cross-patch connections, SPA maximizes th e signal processing for both local features and global contextual representation s. The proposed architecture generates a state-of-the-art performance of 0.29 RM SE for the reconstruction of PPG to ECG on the BIDMC database, surpassing prior studies. We also evaluated this model on the MIMIC-III dataset, achieving a 95.9 % accuracy in CVD detection, and on the PPG-BP dataset, achieving 75.9% accuracy in related CVD diabetes detection, indicating its generalizability. As a proof of concept, an earring wearable named PEARL (prototype), was designed to scale u p the point-of-care (POC) healthcare system.

**********

Token Pooling in Vision Transformers for Image Classification

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastega ri, Oncel Tuzel; Proceedings of the IEEE/CVF Winter Conference on Applications o f Computer Vision (WACV), 2023, pp. 12-21

Pooling is commonly used to improve the computation-accuracy trade-off of convol utional networks. By aggregating neighboring feature values on the image grid, p ooling layers downsample feature maps while maintaining accuracy. In transformer s, however, tokens are processed individually and do not necessarily lie on regu lar grids. Utilizing pooling methods designed for image grids (e.g., average poo ling) can thus be sub-optimal for transformers, as shown by our experiments. In this paper, we propose Token Pooling to downsample tokens in vision transformers . We take a new perspective --- instead of assuming tokens form a regular grid, we treat them as discrete (and irregular) samples of a continuous signal. Given a target number of tokens, Token Pooling finds the set of tokens that best appro ximates the underlying continuous signal. We rigorously evaluate the proposed me thod on the standard transformer architecture (ViT/DeiT), and our experiments sh ow that Token Pooling significantly improves the computation-accuracy trade-off without any further modifications to the architecture. On ImageNet-1k, Token Poo ling enables DeiT-Ti to achieve the same top-1 accuracy while using 42% fewer co mputations.

**********

Lightweight Network for Video Motion Magnification

Jasdeep Singh, Subrahmanyam Murala, G. Sankara Raju Kosuru; Proceedings of the I EEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2 041-2050

Video motion magnification provides information to understand the subtle changes present in objects for applications like industrial, healthcare, sports, etc. M ost state-ofthe-art (SOTA) methods use hand-crafted bandpass filters, which requ ire prior information for the motion magnification, produces ringing artifacts, and small magnification in dynamic scenarios etc. While others use deep-learning based techniques, but their output suffers from artificially induced motion, di stortions, blurriness, etc. Further, SOTA methods are computationally complex, w hich makes them less suitable for real-time applications. To address these probl ems, we proposed deep learning based simple yet effective solution for motion ma gnification. The proposed method uses a feature sharing and appearance encoder f or better motion magnification with less distortions, artifacts etc. Additionall

y, for reducing magnification of noise and other unwanted changes, proxy-model b
ased training is proposed. A computationally lightweight model (  0.12 M paramet
ers) is proposed along with the base model. The performance of the proposed mode
ls is tested qualitatively and quantitatively, with the SOTA methods. Results de
monstrate the effectiveness of the proposed lightweight and base model over the
existing SOTA methods.
**********

Fantastic Style Channels and Where To Find Them: A Submodular Framework for Disc
overing Diverse Directions in GANs
Enis Simsar, Umut Kocasari, Ezgi Gülperi Er, Pinar Yanardag; Proceedings of the
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp.
4731-4740
The discovery of interpretable directions in the latent spaces of pre-trained GA
N models has recently become a popular topic. In particular, StyleGAN2 has enabl
ed various image generation and manipulation tasks due to its rich and disentang
led latent spaces. However, the discovery of such directions is typically made e
ither in a supervised manner, which requires annotated data for each desired man
ipulation, or in an unsupervised manner, which requires a manual effort to ident
ify the directions. As a result, existing work typically finds only a handful of
 directions in which controllable edits can be made. In this study, we design a
novel submodular framework that finds the most representative and diverse subset
 of directions in the latent space of StyleGAN2. Our approach takes advantage of
 the latent space of channel-wise style parameters, so-called stylespace, in whi
ch we cluster channels that perform similar manipulations into groups. Our frame
work promotes diversity by using the notion of clusters and can be efficiently s
olved with a greedy optimization scheme. We evaluate our framework with qualitat
ive and quantitative experiments and show that our method finds more diverse and
 disentangled directions.
**********

Improving Pixel-Level Contrastive Learning by Leveraging Exogenous Depth Informa
tion
Ahmed Ben Saad, Kristina Prokopetc, Josselin Kherroubi, Axel Davy, Adrien Courto
is, Gabriele Facciolo; Proceedings of the IEEE/CVF Winter Conference on Applicat
ions of Computer Vision (WACV), 2023, pp. 2380-2389
Self-supervised representation learning based on Contrastive Learning (CL) has b
een the subject of much attention in recent years. This is due to the excellent
results obtained on a variety of subsequent tasks (in particular classification)
, without requiring a large amount of labeled samples. However, most reference C
L algorithms (such as SimCLR and MoCo, but also BYOL and Barlow Twins) are not a
dapted to pixel-level downstream tasks. One existing solution known as PixPro pr
oposes a pixel-level approach that is based on filtering of pairs of positive/ne
gative image crops of the same image using the distance between the crops in the
 whole image. We argue that this idea can be further enhanced by incorporating s
emantic information provided by exogenous data as an additional selection filter
, which can be used (at training time) to improve the selection of the pixel-lev
el positive/negative samples. In this paper we will focus on the depth informati
on, which can be obtained by using a depth estimation network or measured from a
vailable data (stereovision, parallax motion, lidar, ...). Scene depth can provi
de meaningful cues to distinguish pixels belonging to different objects based on
 their depth. We show that using this exogenous information in the contrastive l
oss leads to improved results and that the learned representations better follow
 the shapes of objects. In addition, we introduce a multi-scale loss that allevi
ates the issue of finding the training parameters adapted to different object si
zes. We demonstrate the effectiveness of our ideas on the Breakout Segmentation
on Borehole Images where we achieve an improvement of 1.9% over PixPro and nearl
y 5% over the supervised baseline. We further validate our technique on the indo
or scene segmentation tasks with ScanNet and outdoor scenes with CityScapes ( 1.
6% and 1.1% improvement over PixPro respectively).
**********

Cooperative Self-Training for Multi-Target Adaptive Semantic Segmentation

Yangsong Zhang, Subhankar Roy, Hongtao Lu, Elisa Ricci, Stéphane Lathuilière; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5604-5613

In this work we address multi-target domain adaptation (MTDA) in semantic segmentation, which consists in adapting a single model from an annotated source dataset to multiple unannotated target datasets that differ in their underlying data distributions. To address MTDA, we propose a self-training strategy that employs pseudo-labels to induce cooperation among multiple domain-specific classifiers. We employ feature stylization as an efficient way to generate image views that forms an integral part of self-training. Additionally, to prevent the network from overfitting to noisy pseudo-labels, we devise a rectification strategy that leverages the predictions from different classifiers to estimate the quality of pseudo-labels. Our extensive experiments on numerous settings, based on four different semantic segmentation datasets, validates the effectiveness of the proposed self-training strategy and shows that our method outperforms state-of-the-art MTDA approaches.

**********

CORL: Compositional Representation Learning for Few-Shot Classification

Ju He, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3890-3899

Few-shot image classification consists of two consecutive learning processes: 1) In the meta-learning stage, the model acquires a knowledge base from a set of training classes. 2) During meta-testing, the acquired knowledge is used to recognize unseen classes from very few examples. Inspired by the compositional representation of objects in humans, we train a neural network architecture that explicitly represents objects as a dictionary of shared components and their spatial composition. In particular, during meta-learning, we train a knowledge base that consists of a dictionary of component representations and a dictionary of component activation maps that encode common spatial activation patterns of components. The elements of both dictionaries are shared among the training classes. During meta-testing, the representation of unseen classes is learned using the component representations and the component activation maps from the knowledge base. Finally, an attention mechanism is used to strengthen those components that are most important for each category. We demonstrate the value of our compositional learning framework for a few-shot classification using miniImageNet, tieredImageNet, CIFAR-FS, and FC100, where we achieve comparable performance.

**********

Unsupervised Video Object Segmentation via Prototype Memory Network

Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5924-5934

Unsupervised video object segmentation aims to segment a target object in the video without a ground truth mask in the initial frame. This challenging task requires extracting features for the most salient common objects within a video sequence. This difficulty can be solved by using motion information such as optical flow, but using only the information between adjacent frames results in poor connectivity between distant frames and poor performance. To solve this problem, we propose a novel prototype memory network architecture. The proposed model effectively extracts the RGB and motion information by extracting superpixel-based component prototypes from the input RGB images and optical flow maps. In addition, the model scores the usefulness of the component prototypes in each frame based on a self-learning algorithm and adaptively stores the most useful prototypes in memory and discards obsolete prototypes. We use the prototypes in the memory bank to predict the next query frame's mask, which enhances the association between distant frames to help with accurate mask prediction. Our method is evaluated on three datasets, achieving state-of-the-art performance. We prove the effectiveness of the proposed model with various ablation studies.

**********

Fine Gaze Redirection Learning With Gaze Hardness-Aware Transformation

Sangjin Park, Daeha Kim, Byung Cheol Song; Proceedings of the IEEE/CVF Winter Co

nference on Applications of Computer Vision (WACV), 2023, pp. 3464-3473
The gaze redirection is a task to adjust the gaze of a given face or eye image t oward the desired direction and aims to learn the gaze direction of a face image through a neural network-based generator. Considering that the prior arts have learned coarse gaze directions, learning fine gaze directions is very challengin g. In addition, explicit discriminative learning of high-dimensional gaze featur es has not been reported yet. This paper presents solutions to overcome the abov e limitations. First, we propose the featurelevel transformation which provides gaze features corresponding to various gaze directions in the latent feature spa ce. Second, we propose a novel loss function for discriminative learning of gaze features. Specifically, features with insignificant or irrelevant effects on ga ze (e.g., head pose and appearance) are set as negative pairs, and important gaz e features are set as positive pairs, and then pair-wise similarity learning is performed. As a result, the proposed method showed a redirection error of only 2 deg for the GazeCapture dataset. This is a 10% better performance than a state-o f-the-art method, i.e., STED. Additionally, the rationale for why latent feature s of various attributes should be discriminated is presented through activation visualization. Code is available at https://github.com/san9569/Gaze-Redir-Learni ng.
**********

Separating Partially-Polarized Diffuse and Specular Reflection Components Under Unpolarized Light Sources
Soma Kajiyama, Taihe Piao, Ryo Kawahara, Takahiro Okabe; Proceedings of the IEEE /CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2549 -2558
Separating diffuse and specular reflection components observed on an object surf ace is important for preprocessing of various computer vision techniques. Conven tionally, diffuse-specular separation based on the polarimetric and color clues assumes that the diffuse/specular reflection components are unpolarized/partiall y polarized under unpolarized light sources. However, the diffuse reflection com ponent is partially polarized in fact, because the diffuse reflectance is maxima l when the polarization direction is parallel to the outgoing plane. Accordingly , we propose a method for separating partially-polarized diffuse and specular re flection components on the basis of the polarization reflection model and the di chromatic reflection model. In particular, our method enables us not only to ach ieve diffuse-specular separation but also to estimate the polarimetric propertie s of the object surface from a single color polarization image. We experimentall y confirmed that our method performs better than the method assuming unpolarized diffuse reflection components.
**********

ScanNeRF: A Scalable Benchmark for Neural Radiance Fields
Luca De Luigi, Damiano Bolognini, Federico Domeniconi, Daniele De Gregorio, Matt eo Poggi, Luigi Di Stefano; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 816-825
In this paper, we propose the first-ever real benchmark thought for evaluating N eural Radiance Fields (NeRFs) and, in general, Neural Rendering (NR) frameworks. We design and implement an effective pipeline for scanning real objects in quan tity and effortlessly. Our scan station is built with less than 500 hardware bud get and can collect roughly 4000 images of a scanned object in just 5 minutes. S uch a platform is used to build ScanNeRF, a dataset characterized by several tra in/val/test splits aimed at benchmarking the performance of modern NeRF methods under different conditions. Accordingly, we evaluate three cutting-edge NeRF var iants on it to highlight their strengths and weaknesses. The dataset is availabl e on our project page, together with an online benchmark to foster the developme nt of better and better NeRFs.
**********

Adaptive Sample Selection for Robust Learning Under Label Noise
Deep Patel, P. S. Sastry; Proceedings of the IEEE/CVF Winter Conference on Appli cations of Computer Vision (WACV), 2023, pp. 3932-3942
Deep Neural Networks (DNNs) have been shown to be susceptible to memorization or

overfitting in the presence of noisily-labelled data. For the problem of robust learning under such noisy data, several algorithms have been proposed. A prominent class of algorithms rely on sample selection strategies wherein, essentially, a fraction of samples with loss values below a certain threshold are selected for training. These algorithms are sensitive to such thresholds, and it is difficult to fix or learn these thresholds. Often, these algorithms also require information such as label noise rates which are typically unavailable in practice. In this paper, we propose an adaptive sample selection strategy that relies only on batch statistics of a given mini-batch to provide robustness against label noise. The algorithm does not have any additional hyperparameters for sample selection, does not need any information on noise rates and does not need access to separate data with clean labels. We empirically demonstrate the effectiveness of our algorithm on benchmark datasets.

**********

Patch-Based Privacy Preserving Neural Network for Vision Tasks

Mitsuhiro Mabuchi, Tetsuya Ishikawa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1550-1559

As machine learning technology is increasingly adopted into a variety of application domains, the potential risks of data leakage are becoming more serious in certain cases where the data contains highly sensitive information. While some privacy-preserving learning mechanisms for image data, such as SplitNN, enable the training of models without sharing private data on a central server, there exists a trade-off between security and computational cost to a client device. We propose a new mechanism to achieve higher level security and lower computational cost on a client device while maintaining model performance. Our approach, called Patch SplitNN, is based on SplitNN architecture that divides a CNN into two networks, called upper and lower. The difference from that previous work is to input individual image patches into multiple upper models, before concatenating their outputs before the lower model. For further improvement of the upper model training, we introduce an additional network and a loss function into the training process. We demonstrate our Patch SplitNN can classify images as accurately as a ResNet18 on various image classification datasets (CIFAR-10, CIFAR-100, and PCam) under multiple conditions (e.g. patching patterns, dropping patches).

**********

Image-Text Pre-Training for Logo Recognition

Mark Hubenthal, Suren Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1145-1154

Open-set logo recognition is commonly solved by first detecting possible logo regions and then matching the detected parts against an ever-evolving dataset of cropped logo images. The matching model, a metric learning problem, is especially challenging for logo recognition due to the mixture of text and symbols in logos. We propose two novel contributions to improve the matching model's performance: (a) using image-text paired samples for pre-training, and (b) an improved metric learning loss function. A standard paradigm of fine-tuning ImageNet pre-trained models fails to discover the text sensitivity necessary to solve the matching problem effectively. This work demonstrates the importance of pre-training on image-text pairs, which significantly improves the performance of a visual embedder trained for the logo retrieval task, especially for more text-dominant classes. We construct a composite public logo dataset combining LogoDet3K, OpenLogo, and FlickrLogos-47 deemed OpenLogoDet3K47. We show that the same vision backbone pre-trained on image-text data, when fine-tuned on OpenLogoDet3K47, achieves 98.6% recall@1, significantly improving performance over pre-training on Imagenet1K (97.6%). We generalize the ProxyNCA++ loss function to propose ProxyNCAHN++ which incorporates class-specific hard negative images. The proposed method sets new state-of-the-art on five public logo datasets considered, with a 3.5% zero-shot recall@1 improvement on LogoDet3K test, 4% on OpenLogo, 6.5% on FlickrLogos-47, 6.2% on Logos In The Wild, and 0.6% on BelgaLogo.

**********

LRA&LDRA: Rethinking Residual Predictions for Efficient Shadow Detection and Removal

Mehmet Kerim Yücel, Valia Dimaridou, Bruno Manganelli, Mete Ozay, Anastasios Drosou, Albert Saà-Garriga; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4925-4935
The majority of the state-of-the-art shadow removal models (SRMs) reconstruct whole input images, where their capacity is needlessly spent on reconstructing non-shadow regions. SRMs that predict residuals remedy this up to a degree, but fall short of providing an accurate and flexible solution. In this paper, we rethink residual predictions and propose Learnable Residual Attention (LRA) and Learnable Dense Reconstruction Attention (LDRA) modules, which operate over the input and the output of SRMs. These modules guide an SRM to concentrate on shadow region reconstruction, and limit reconstruction of non-shadow regions. The modules improve shadow removal (up to 20%) and detection accuracy across various backbones, and even improve the accuracy of other removal methods (up to 10%). In addition, the modules have minimal overhead (+<1MB memory) and are implemented in a few lines of code. Furthermore, to combat the challenge of training SRMs with small datasets, we present a synthetic dataset generation pipeline. Using our pipeline, we create a dataset called PITSA, which has 10 times more unique shadow-free images than the largest benchmark dataset. Pre-training models on the PITSA significantly improves shadow removal (+2 MAE on shadow regions) and detection accuracy of multiple methods. Our results show that LRA&LDRA, when plugged into a lightweight architecture pre-trained on the PITSA, outperform state-of-the-art shadow removal (+0.7 all-region MAE) and detection (+0.1 BER) methods on the benchmark ISTD and SRD datasets, despite running faster (+5%) and consuming less memory (x150).
**********

Urban Scene Semantic Segmentation With Low-Cost Coarse Annotation
Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, Bernt Schiele; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5978-5987
For best performance, today's semantic segmentation methods use large and carefully labeled datasets, requiring expensive annotation budgets. In this work, we show that coarse annotation is a low-cost but highly effective alternative for training semantic segmentation models. Considering the urban scene segmentation scenario, we leverage cheap coarse annotations for real-world captured data, as well as synthetic data to train our model and show competitive performance compared with fully annotated real-world data. Specifically, we propose a coarse-to fine self-training framework that generates pseudo labels for unlabeled regions of the coarsely annotated data, using synthetic data to improve predictions around the boundaries between semantic classes, and using cross-domain data augmentation to increase diversity. Our extensive experimental results on Cityscapes and BDD100k datasets demonstrate that our method achieves a significantly better performance vs annotation cost tradeoff, yielding a comparable performance to fully annotated data with only a small fraction of the annotation budget. Also, when used as pretraining, our framework performs better compared to the standard fully supervised setting.
**********

TransVLAD: Multi-Scale Attention-Based Global Descriptors for Visual Geo-Localization
Yifan Xu, Pourya Shamsolmoali, Eric Granger, Claire Nicodeme, Laurent Gardes, Jie Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2840-2849
Visual geo-localization remains a challenging task due to variations in the appearance and perspective among captured images. This paper introduces an efficient TransVLAD module, which aggregates attention-based feature maps into a discriminative and compact global descriptor. Unlike existing methods that generate feature maps using only convolutional neural networks (CNNs), we propose a sparse transformer to encode global dependencies and compute attention-based feature maps, which effectively reduces visual ambiguities that occurs in large-scale geo-localization problems. A positional embedding mechanism is used to learn the corresponding geometric configurations between query and gallery images. A grouped VL

AD layer is also introduced to reduce the number of parameters, and thus construct an efficient module. Finally, rather than only learning from the global descriptors on entire images, we propose a self-supervised learning method to further encode more information from multi-scale patches between the query and positive gallery images. Extensive experiments on three challenging large-scale datasets indicate that our model outperforms state-of-the-art models, and has lower computational complexity.
**********
Ancestor Search: Generalized Open Set Recognition via Hyperbolic Side Information Learning
Xiwen Dengxiong, Yu Kong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4003-4012
Different from the open set recognition, generalized open set recognition learns the most similar known classes for unseen samples using known classes samples and side information of known classes. It is challenging because hierarchically structured side information is distorted when features are embedded in the Euclidean space in existing literature, which incurs the difficulty of identifying the unseen samples. In this paper, we introduce a side information learning algorithm for generalized open set recognition based on the hyperbolic space to alleviate the distortion and accurately identify the unknown samples. Specifically, we propose a hyperbolic side information learning framework to identify the unseen samples and an ancestor search algorithm to search the most similar ancestor from the taxonomy of selected known classes. Experiments on CUB-200 and AWA 2 datasets show that our method improves the performance of generalized open set recognition by a large margin.
**********
Unsupervised 4D LiDAR Moving Object Segmentation in Stationary Settings With Multivariate Occupancy Time Series
Thomas Kreutz, Max Mühlhäuser, Alejandro Sanchez Guinea; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1644-1653
In this work, we address the problem of unsupervised moving object segmentation (MOS) in 4D LiDAR data recorded from a stationary sensor, where no ground truth annotations are involved. Deep learning-based state-of-the-art methods for LiDAR MOS strongly depend on annotated ground truth data, which is expensive to obtain and scarce in existence. To close this gap in the stationary setting, we propose a novel 4D LiDAR representation based on multivariate time series that relaxes the problem of unsupervised MOS to a time series clustering problem. More specifically, we propose modeling the change in occupancy of a voxel by a multivariate occupancy time series (MOTS), which captures spatio-temporal occupancy changes on the voxel level and its surrounding neighborhood. To perform unsupervised MOS, we train a neural network in a self-supervised manner to encode MOTS into voxel-level feature representations, which can be partitioned by a clustering algorithm into moving or stationary. Experiments on stationary scenes from the Raw KITTI dataset show that our fully unsupervised approach achieves performance that is comparable to that of supervised state-of-the-art approaches.
**********
Modeling the Lighting in Scenes As Style for Auto White-Balance Correction
Furkan K■nl■, Do■a Y■lmaz, Bar■■ Özcan, Furkan K■raç; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4903-4913
Style may refer to different concepts (e.g. painting style, hairstyle, texture, color, filter, etc.) depending on how the feature space is formed. In this work, we propose a novel idea of interpreting the lighting in the single- and multi-illuminant scenes as the concept of style. To verify this idea, we introduce an enhanced auto white-balance (AWB) method that models the lighting in single- and mixed-illuminant scenes as the style factor. Our AWB method does not require any illumination estimation step, yet contains a network learning to generate the weighting maps of the images with different WB settings. Proposed network utilizes the style information, extracted from the scene by a multi-head style extracti

on module. AWB correction is completed after blending these weighting maps and t
he scene. Experiments on single- and mixed-illuminant datasets demonstrate that
our proposed method achieves promising correction results when compared to the r
ecent works. This shows that the lighting in the scenes with multiple illuminati
ons can be modeled by the concept of style. Source code and trained models are a
vailable on https://github.com/birdortyedi/lighting-as-style-awb-correction.
**********

Semantic Segmentation in Aerial Imagery Using Multi-Level Contrastive Learning W
ith Local Consistency

Maofeng Tang, Konstantinos Georgiou, Hairong Qi, Cody Champion, Marc Bosch; Proc
eedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WA
CV), 2023, pp. 3798-3807

Semantic segmentation in large-scale aerial images is an extremely challenging t
ask. On one hand, the limited ground truth, as compared to the vast area the ima
ges cover, greatly hinders the development of supervised representation learning
. On the other hand, the large footprint from remote sensing raises new challeng
es for semantic segmentation. In addition, the complex and ever changing image a
cquisition conditions further complicate the problem where domain shifting commo
nly occurs. In this paper, we exploit self-supervised contrastive learning (CL)
methodologies for semantic segmentation in aerial imagery. In addition to perfor
ming CL at the feature level as most practices do, we add another level of contr
astive learning, at the semantic level, taking advantage of the segmentation out
put from the downstream task. Further, we embed local mutual information in the
semantic-level CL to enforce local consistency. This has largely enhanced the re
presentation power at each pixel and improved the generalization capacity of the
 trained model. We refer to the proposed approach as multi-level contrastive lea
rning with local consistency (mCL-LC). The experimental results on different ben
chmarks indicate that the proposed mCL-LC exhibits superior performance as compa
red to other state-of-the-art contrastive learning frameworks for the semantic s
egmentation task. mCL-LC also carries better generalization capacity especially
when domain shifting exists.
**********

Fast Differentiable Transient Rendering for Non-Line-of-Sight Reconstruction

Markus Plack, Clara Callenberg, Monika Schneider, Matthias B. Hullin; Proceeding
s of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2
023, pp. 3067-3076

Research into non-line-of-sight imaging problems has gained momentum in recent y
ears motivated by intriguing prospective applications in e.g. medicine and auton
omous driving. While transient image formation is well understood and there exis
t various reconstruction approaches for non-line-of-sight scenes that combine ef
ficient forward renderers with optimization schemes, those approaches suffer fro
m runtimes in the order of hours even for moderately sized scenes. Furthermore,
the ill-posedness of the inverse problem often leads to instabilities in the opt
imization. Inspired by the latest advances in direct-line-of-sight inverse rende
ring that have led to stunning results for reconstructing scene geometry and app
earance, we present a fast differentiable transient renderer that accelerates th
e inverse rendering runtime to minutes on consumer hardware, making it possible
to apply inverse transient imaging on a wider range of tasks and in more time-cr
itical scenarios. We demonstrate its effectiveness on a series of applications u
sing various datasets and show that it can be used for self-supervised learning.
**********

Neural Distributed Image Compression With Cross-Attention Feature Alignment

Nitish Mital, Ezgi Özyilkan, Ali Garjani, Deniz Gündüz; Proceedings of the IEEE/
CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2498-
2507

We consider the problem of compressing an information source when a correlated o
ne is available as side information only at the decoder side, which is a special
 case of the distributed source coding problem in information theory. In particu
lar, we consider a pair of stereo images, which have overlapping fields of view,
 and are captured by a synchronized and calibrated pair of cameras as correlated

image sources. In previously proposed methods, the encoder transforms the input image to a latent representation using a deep neural network, and compresses the quantized latent representation losslessly using entropy coding. The decoder decodes the entropy-coded quantized latent representation, and reconstructs the input image using this representation and the available side information. In the proposed method, the decoder employs a cross-attention module to align the feature maps obtained from the received latent representation of the input image and a latent representation of the side information. We argue that aligning the correlated patches in the feature maps allows better utilization of the side information. We empirically demonstrate the competitiveness of the proposed algorithm on KITTI and Cityscape datasets of stereo image pairs. Our experimental results show that the proposed architecture is able to exploit the decoder-only side information in a more efficient manner compared to previous works.

**********

Burst Vision Using Single-Photon Cameras

Sizhuo Ma, Paul Mos, Edoardo Charbon, Mohit Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5375-5385

Single-photon avalanche diodes (SPADs) are novel image sensors that record the arrival of individual photons at extremely high temporal resolution. In the past, they were only available as single pixels or small-format arrays, for various active imaging applications such as LiDAR and microscopy. Recently, high-resolution SPAD arrays up to 3.2 megapixel have been realized, which for the first time may be able to capture sufficient spatial details for general computer vision tasks, purely as a passive sensor. However, existing vision algorithms are not directly applicable on the binary data captured by SPADs. In this paper, we propose developing quanta vision algorithms based on burst processing for extracting scene information from SPAD photon streams. With extensive real-world data, we demonstrate that current SPAD arrays, along with burst processing as an example plug-and-play algorithm, are capable of a wide range of downstream vision tasks in extremely challenging imaging conditions including fast motion, low light (<5 lux) and high dynamic range. To our knowledge, this is the first attempt to demonstrate the capabilities of SPAD sensors for a wide gamut of real-world computer vision tasks including object detection, pose estimation, SLAM, and text recognition. We hope this work will inspire future research into developing computer vision algorithms for robust scene inference in extreme scenarios using single-photon cameras.

**********

GlobalFlowNet: Video Stabilization Using Deep Distilled Global Motion Estimates

Jerin Geo James, Devansh Jain, Ajit Rajwade; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5078-5087

Videos shot by laymen using hand-held cameras contain undesirable shaky motion. Estimating the global motion between successive frames, in a manner not influenced by moving objects, is central to many video stabilization techniques, but poses significant challenges. A large body of work uses 2D affine transformations or homography for the global motion. However, in this work, we introduce a more general representation scheme, which adapts any existing optical flow network to ignore the moving objects and obtain a spatially smooth approximation of the global motion between video frames. We achieve this by a knowledge distillation approach, where we first introduce a low pass filter module into the optical flow network to constrain the predicted optical flow to be spatially smooth. This becomes our student network, named as GLOBALFLOWNET. Then, using the original optical flow network as the teacher network, we train the student network using a robust loss function. Given a trained GLOBALFLOWNET, we stabilize videos using a two stage process. In the first stage, we correct the instability in affine parameters using a quadratic programming approach constrained by a user-specified cropping limit to control loss of field of view. In the second stage, we stabilize the video further by smoothing global motion parameters, expressed using small number of discrete cosine transform coefficients. In extensive experiments on a variety of different videos, our technique outperforms state of the art techniques in terms of subjective quality and different quantitative measures of video stab

ility. Additionally, we present a new measure for evaluation of video stabilizat ion based on the flow generated by GLOBALFLOWNET and argue that it is based on a more general motion model in contrast to the affine motion model on which most existing measures are based. The source code is publicly available at https://gi thub.com/GlobalFlowNet/GlobalFlowNet

**********

Action-Aware Masking Network With Group-Based Attention for Temporal Action Loca lization

Tae-Kyung Kang, Gun-Hee Lee, Kyung-Min Jin, Seong-Whan Lee; Proceedings of the I EEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6 058-6067

Temporal Action Localization (TAL) is a significant and challenging task that se arches for subtle human activities in an untrimmed video. To extract snippet-lev el video features, existing TAL methods commonly use video encoders pre-trained on short-video classification datasets. However, the snippet-level features can incur ambiguity between consecutive frames due to short and poor temporal inform ation, disrupting the precise prediction of action instances. Several methods in corporating temporal relations have been proposed to mitigate this problem; howe ver, they still suffer from poor video features. To address this issue, we propo se a novel temporal action localization framework called an Action-aware Masking Network (AMNet). Our method simultaneously refines video features using action- aware attention and considers inherent temporal relations using self-attention a nd cross-attention mechanisms. First, we present an Action Masking Encoder (AME) that generates an action-aware mask to represent positive characteristics, whic h is then used to refine snippet-level features to be more salient around action s. Second, we design a Group Attention Module (GAM), which models relations of t emporal information and exchanges mutual information by dividing the features in to two groups, i.e., long and short-groups. Extensive experiments and ablation s tudies on two primary benchmark datasets demonstrate the effectiveness of AMNet, and our method achieves state-of-the-art performances on THUMOS-14 and Activity Net1.3.

**********

A Deep Neural Framework To Detect Individual Advertisement (Ad) From Videos

Zongyi Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Com puter Vision (WACV), 2023, pp. 3578-3587

Detecting commercial Ads from a video is important. For example, the commercial break frequency and duration are two metrics to measure the user experience for streaming service providers such as Amazon IMDb TV. The detection can be done in trusively by intercepting the network traffic and then parsing the service provi ders data and logs, or non-intrusively by capturing the videos streamed by conte nt providers and then analyzing using the computer vision technologies. In this paper, we present a non-intrusive framework that is able to not only detect an A d section, but also segment out individual Ads. We show that our algorithm is no t only scalable because it uses light weight audio data to do global segmentatio n, but also robust as the Ad classifier is able to handle different types of con tents captured from the popular streaming services such as the IMDb TV, Hulu, Cr ackleTV, and Prime Video (PV) live sports.

**********

Normality Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection

Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2665-2674

Weakly supervised Video Anomaly Detection (wVAD) aims to distinguish anomalies f rom normal events based on video-level supervision. Most existing works utilize Multiple Instance Learning (MIL) with ranking loss to tackle this task. These me thods, however, rely on noisy predictions from a MIL-based classifier for target instance selection in ranking loss, degrading model performance. To overcome th is problem, we propose Normality Guided Multiple Instance Learning (NG-MIL) fram ework, which encodes diverse normal patterns from noise-free normal videos into

prototypes for constructing a similarity-based classifier. By ensembling predictions of two classifiers, our method could refine the anomaly scores, reducing training instability from weak labels. Moreover, we introduce normality clustering and normality guided triplet loss constraining inner bag instances to boost the effect of NG-MIL and increase the discriminability of classifiers. Extensive experiments on three public datasets (ShanghaiTech, UCF-Crime, XD-Violence) demonstrate that our method is comparable to or better than existing weakly supervised methods, achieving state-of-the-art results.
**********

ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-Based Mixing
Giulio Mattolin, Luca Zanella, Elisa Ricci, Yiming Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 423-433
Unsupervised Domain Adaptation (UDA) for object detection aims to adapt a model trained on a source domain to detect instances from a new target domain for which annotations are not available. Different from traditional approaches, we propose ConfMix, the first method that introduces a sample mixing strategy based on region-level detection confidence for adaptive object detector learning. We mix the local region of the target sample that corresponds to the most confident pseudo detections with a source image, and apply an additional consistency loss term to gradually adapt towards the target data distribution. In order to robustly define a confidence score for a region, we exploit the confidence score per pseudo detection that accounts for both the detector-dependent confidence and the bounding box uncertainty. Moreover, we propose a novel pseudo labelling scheme that progressively filters the pseudo target detections using the confidence metric that varies from a loose to strict manner along the training. We perform extensive experiments with three datasets, achieving state-of-the-art performance in two of them and approaching the supervised target model performance in the other. Code is available at https://github.com/giuliomattolin/ConfMix.
**********

ROMA: Run-Time Object Detection To Maximize Real-Time Accuracy
JunKyu Lee, Blesson Varghese, Hans Vandierendonck; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6405-6414
This paper analyzes the effects of dynamically varying video contents and detection latency on the real-time detection accuracy of a detector and proposes a new run-time accuracy variation model, ROMA, based on the findings from the analysis. ROMA is designed to select an optimal detector out of a set of detectors in real time without label information to maximize real-time object detection accuracy. ROMA utilizing four YOLOv4 detectors on an NVIDIA Jetson Nano shows real-time accuracy improvements by 4 to 37% for a scenario of dynamically varying video contents and detection latency consisting of MOT17Det and MOT20Det datasets, compared to individual YOLOv4 detectors and two state-of-the-art runtime techniques.
**********

PointNeuron: 3D Neuron Reconstruction via Geometry and Topology Learning of Point Clouds
Runkai Zhao, Heng Wang, Chaoyi Zhang, Weidong Cai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5787-5797
Digital neuron reconstruction from 3D microscopy images is an essential technique for investigating brain connectomics and neuron morphology. Existing reconstruction frameworks use convolution-based segmentation networks to partition the neuron from noisy backgrounds before applying the tracing algorithm. The tracing results are sensitive to the raw image quality and segmentation accuracy. In this paper, we propose a novel framework for 3D neuron reconstruction. Our key idea is to use the geometric representation power of the point cloud to better explore the intrinsic structural information of neurons. Our proposed framework adopts one graph convolutional network to predict the neural skeleton points and another one to produce the connectivity of these points. We finally generate the target SWC file through the interpretation of the predicted point coordinates, radiu

s, and connections. Evaluated on the Janelia-Fly dataset from the BigNeuron project, we show that our framework achieves competitive neuron reconstruction performance. Our geometry and topology learning of point clouds could further benefit 3D medical image analysis, such as cardiac surface reconstruction. Our code is available at https://github.com/RunkaiZhao/PointNeuron.

**********

High-Resolution Depth Estimation for 360deg Panoramas Through Perspective and Panoramic Depth Images Registration

Chi-Han Peng, Jiayao Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3116-3125

We propose a novel approach to compute high-resolution (2048x1024 and higher) depths for panoramas that is significantly faster and qualitatively and qualitatively more accurate than the current state-of-the-art method (360MonoDepth). As traditional neural network-based methods have limitations in the output image sizes (up to 1024x512) due to GPU memory constraints, both 360MonoDepth and our method rely on stitching multiple perspective disparity or depth images to come out a unified panoramic depth map. However, to achieve globally consistent stitching, 360MonoDepth relied on solving extensive disparity map alignment and Poisson-based blending problems, leading to high computation time. Instead, we propose to use an existing panoramic depth map (computed in real-time by any panorama-based method) as the common target for the individual perspective depth maps to register to. This key idea made producing globally consistent stitching results from a straightforward task. Our experiments show that our method generates qualitatively better results than existing panorama-based methods, and further outperforms them quantitatively on datasets unseen by these methods.

**********

A Suspect Identification Framework Using Contrastive Relevance Feedback

Devansh Gupta, Aditya Saini, Sarthak Bhagat, Shagun Uppal, Rishi Raj Jain, Drishti Bhasin, Ponnurangam Kumaraguru, Rajiv Ratn Shah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4361-4369

Suspect Identification is one of the most pivotal aspects of a forensic and criminal investigation. A significant amount of time and skill is devoted to creating sketches for it and requires a fair amount of recollections from the witness to provide a useful sketch. We devise a method that aims to automate the process of suspect identification and model this problem by iteratively retrieving images from feedback provided by the user. Compared to standard image retrieval tasks, interactive facial image retrieval is specifically more challenging due to the high subjectivity involved in describing a person's facial attributes and appropriately evolving with the preferences put forward by the user. Our method uses a relatively simpler form of supervision by utilizing the user's feedback to label images as either similar or dissimilar to their mental image of the suspect based on which we propose a loss function using the contrastive learning paradigm that is optimized in an online fashion. We validate the efficacy of our proposed approach using a carefully designed testbed to simulate user feedback and a large-scale user study. We empirically show that our method iteratively improves personalization, leading to faster convergence and enhanced recommendation relevance, thereby, improving user satisfaction. Our proposed framework is being designed for real-time use in the metropolitan crime investigation department, and thus is also equipped with a user-friendly web interface with a real-time experience for suspect retrieval.

**********

Compressing Explicit Voxel Grid Representations: Fast NeRFs Become Also Small

Chenxi Lola Deng, Enzo Tartaglione; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1236-1245

NeRFs have revolutionized the world of per-scene radiance field reconstruction because of their intrinsic compactness. One of the main limitations of NeRFs is their slow rendering speed, both at training and inference time. Recent research addressing this issue focuses on optimisation of an explicit voxel grid (EVG) that represents the scene, which can be paired with neural networks to learn radiance fields. This approach significantly enhances the speed both at train and inf

erence time, but at the cost of large memory occupation. In this work we propose Re:NeRF, an approach specifically designed for targeting EVG-NeRFs compressibility, which aims to reduce memory storage of NeRF models while maintaining comparable performance. We benchmark our approach with three different EVG-NeRF architectures on three popular benchmarks, showing Re:NeRF's broad usability and effectiveness.
**********

Treating Motion as Option To Reduce Motion Dependency in Unsupervised Video Object Segmentation
Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Chaewon Park, Donghyeong Kim, Sangyoun Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5140-5149
Unsupervised video object segmentation (VOS) aims to detect the most salient object in a video sequence at the pixel level. In unsupervised VOS, most state-of-the-art methods leverage motion cues obtained from optical flow maps in addition to appearance cues to exploit the property that salient objects usually have distinctive movements compared to the background. However, as they are overly dependent on motion cues, which may be unreliable in some cases, they cannot achieve stable prediction. To reduce this motion dependency of existing two-stream VOS methods, we propose a novel motion-as-option network that optionally utilizes motion cues. Additionally, to fully exploit the property of the proposed network that motion is not always required, we introduce a collaborative network learning strategy. On all the public benchmark datasets, our proposed network affords state-of-the-art performance with real-time inference speed.
**********

Sim2real Transfer Learning for Point Cloud Segmentation: An Industrial Application Case on Autonomous Disassembly
Chengzhi Wu, Xuelei Bi, Julius Pfrommer, Alexander Cebulla, Simon Mangold, Jürgen Beyerer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4531-4540
On robotics computer vision tasks, generating and annotating large amounts of data from real-world for the use of deep learning-based approaches is often difficult or even impossible. A common strategy for solving this problem is to apply simulation-to-reality (sim2real) approaches with the help of simulated scenes. While the majority of current robotics vision sim2real work focuses on image data, we present an industrial application case that uses sim2real transfer learning for point cloud data. We provide insights on how to generate and process synthetic point cloud data in order to achieve better performance when the learned model is transferred to real-world data. The issue of imbalanced learning is investigated using multiple strategies. A novel patch-based attention network is proposed additionally to tackle this problem.
**********

Vis2Rec: A Large-Scale Visual Dataset for Visit Recommendation
Michaël Soumm, Adrian Popescu, Bertrand Delezoide; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2987-2997
Most recommendation datasets for tourism are restricted to one world region and rely on explicit data such as check-ins. However, in reality, tourists visit various places worldwide and document their trips primarily through photos. These images contain a wealth of raw information that can be used to capture users' preferences and recommend personalized content. Visual content was already used in past works, but no large-scale publicly-available dataset that gives access to users' personal images exists for recommender systems. As such a resource would open-up possibilities for new image-based recommendation algorithms, we introduce Vis2Rec, a new dataset based on visit data extracted from users' Flickr photographic streams, which includes over 7 million photos, 36k recognizable points of interest, and 14k user profiles. Google Landmarks v2 is used as an auxiliary dataset to identify points of interest in users' photos, using a state-of-the-art image-matching deep architecture. Image-based user profiles are then constituted by aggregating the points of interest detected for each user. In addition, ground truth visits were determined for the test subset in order to enable accurate e

valuation. Finally, we benchmark Vis2Rec using various existing recommender systems, and discuss the possibilities opened up by the availability of user images, as well as the societal issues that come with them. Following good practice in dataset sharing, Vis2Rec is created using only freely distributable content, and additional anonymization is performed to ensure the privacy of users. The raw dataset and the preprocessed user profiles will be publicly available at https://github.com/MSoumm/Vis2Rec.
**********

3D Neural Sculpting (3DNS): Editing Neural Signed Distance Functions
Petros Tzathas, Petros Maragos, Anastasios Roussos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4521-4530
In recent years, implicit surface representations through neural networks that encode the signed distance have gained popularity and have achieved state-of-the-art results in various tasks (e.g. shape representation, shape reconstruction and learning shape priors). However, in contrast to conventional shape representations such as polygon meshes, the implicit representations cannot be easily edited and existing works that attempt to address this problem are extremely limited. In this work, we propose the first method for efficient interactive editing of signed distance functions expressed through neural networks, allowing free-form editing. Inspired by 3D sculpting software for meshes, we use a brush-based framework that is intuitive and can in the future be used by sculptors and digital artists. In order to localize the desired surface deformations, we regulate the network by using a copy of it to sample the previously expressed surface. We introduce a novel framework for simulating sculpting-style surface edits, in conjunction with interactive surface sampling and efficient adaptation of network weights. We qualitatively and quantitatively evaluate our method in various different 3D objects and under many different edits. The reported results clearly show that our method yields high accuracy, in terms of achieving the desired edits, while in the same time preserving the geometry outside the interaction areas.
**********

Learning To Detect 3D Lanes by Shape Matching and Embedding
Ruixin Liu, Zhihao Guan, Zejian Yuan, Ao Liu, Tong Zhou, Tang Kun, Erlong Li, Chao Zheng, Shuqi Mei; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4291-4299
3D lane detection based on LiDAR point clouds is a challenging task that requires precise locations, accurate topologies, and distinguishable instances. In this paper, we propose a dual-level shape attention network (DSANet) with two branches for high-precision 3D lane predictions. Specifically, one branch predicts the refined lane segment shapes and the shape embeddings that encode the approximate lane instance shapes, the other branch detects the coarse-grained structures of the lane instances. In the training stage, two-level shape matching loss functions are introduced to jointly optimize the shape parameters of the two-branch outputs, which are simple yet effective for precision enhancement. Furthermore, a shape-guided segments aggregator is proposed to help local lane segments aggregate into complete lane instances, according to the differences of instance shapes predicted at different levels. Experiments conducted on our BEV-3DLanes dataset demonstrate that our method outperforms previous methods.
**********

Holistic Interaction Transformer Network for Action Detection
Gueter Josmy Faure, Min-Hung Chen, Shang-Hong Lai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3340-3350
Actions are about how we interact with the environment, including other people, objects, and ourselves. In this paper, we propose a novel multi-modal Holistic Interaction Transformer Network (HIT) that leverages the largely ignored, but critical hand and pose information essential to most human actions. The proposed HIT network is a comprehensive bi-modal framework that comprises an RGB stream and a pose stream. Each of them separately models person, object, and hand interactions. Within each sub-network, an Intra-Modality Aggregation module (IMA) is introduced that selectively merges individual interaction units. The resulting features from each modality are then glued using an Attentive Fusion Mechanism (AFM)

. Finally, we extract cues from the temporal context to better classify the occu
rring actions using cached memory. Our method significantly outperforms previous
 approaches on the J-HMDB, UCF101-24, and MultiSports datasets. We also achieve
competitive results on AVA. The code will be available at https://github.com/jos
lefaure/HIT.
**********

Keys To Better Image Inpainting: Structure and Texture Go Hand in Hand
Jitesh Jain, Yuqian Zhou, Ning Yu, Humphrey Shi; Proceedings of the IEEE/CVF Win
ter Conference on Applications of Computer Vision (WACV), 2023, pp. 208-217
Deep image inpainting has made impressive progress with recent advances in image
 generation and processing algorithms. We claim that the performance of inpainti
ng algorithms can be better judged by the generated structures and textures. Str
uctures refer to the generated object boundary or novel geometric structures wit
hin the hole, while texture refers to high-frequency details, especially man-mad
e repeating patterns filled inside the structural regions. We believe that bette
r structures are usually obtained from a coarse-to-fine GAN-based generator netw
ork while repeating patterns nowadays can be better modeled using state-of-the-a
rt high-frequency fast fourier convolutional layers. In this paper, we propose a
 novel inpainting network combining the advantages of the two designs. Therefore
, our model achieves a remarkable visual quality to match state-of-the-art perfo
rmance in both structure generation and repeating texture synthesis using a sing
le network. Extensive experiments demonstrate the effectiveness of the method, a
nd our conclusions further highlight the two critical factors of image inpaintin
g quality, structures, and textures, as the future design directions of inpainti
ng networks.
**********

Reconstructing Humpty Dumpty: Multi-Feature Graph Autoencoder for Open Set Actio
n Recognition
Dawei Du, Ameya Shringi, Anthony Hoogs, Christopher Funk; Proceedings of the IEE
E/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 337
1-3380
Most action recognition datasets and algorithms assume a closed world, where all
 test samples are instances of the known classes. In open set problems, test sam
ples may be drawn from either known or unknown classes. Existing open set action
 recognition methods are typically based on extending closed set methods by addi
ng post hoc analysis of classification scores or feature distances and do not ca
pture the relations among all the video clip elements. Our approach uses the rec
onstruction error to determine the novelty of the video since unknown classes ar
e harder to put back together and thus have a higher reconstruction error than v
ideos from known classes. We refer to our solution to the open set action recogn
ition problem as "Humpty Dumpty", due to its reconstruction abilities. Humpty Du
mpty is a novel graph-based autoencoder that accounts for contextual and semanti
c relations among the clip pieces for improved reconstruction. A larger reconstr
uction error leads to an increased likelihood that the action can not be reconst
ructed, i.e., can not put Humpty Dumpty back together again, indicating that the
 action has never been seen before and is novel/unknown. Extensive experiments a
re performed on two publicly available action recognition datasets including HMD
B-51 and UCF-101, showing the state-of-the-art performance for open set action r
ecognition.
**********

Attention Attention Everywhere: Monocular Depth Prediction With Skip Attention
Ashutosh Agarwal, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on
 Applications of Computer Vision (WACV), 2023, pp. 5861-5870
Monocular Depth Estimation (MDE) aims to predict pixel-wise depth given a single
 RGB image. For both, the convolutional as well as the recent attention-based mo
dels, encoder-decoder-based architectures have been found to be useful due to th
e simultaneous requirement of global context and pixel-level resolution. Typical
ly, a skip connection module is used to fuse the encoder and decoder features, w
hich comprises of feature map concatenation followed by a convolution operation.
 Inspired by the demonstrated benefits of attention in a multitude of computer v

ision problems, we propose an attention-based fusion of encoder and decoder features. We pose MDE as a pixel query refinement problem, where coarsest-level encoder features are used to initialize pixel-level queries, which are then refined to higher resolutions by the proposed Skip Attention Module (SAM). We formulate the prediction problem as ordinal regression over the bin centers that discretize the continuous depth range and introduce a Bin Center Predictor (BCP) module that predicts bins at the coarsest level using pixel queries. Apart from the benefit of image adaptive depth binning, the proposed design helps learn improved depth embedding in initial pixel queries via direct supervision from the ground truth. Extensive experiments on the two canonical datasets, NYUV2 and KITTI, show that our architecture outperforms the state-of-the-art by 5.3% and 3.9%, respectively, along with an improved generalization performance by 9.4% on the SUNRGBD dataset.

**********

360MVSNet: Deep Multi-View Stereo Network With 360deg Images for Indoor Scene Reconstruction

Ching-Ya Chiu, Yu-Ting Wu, I-Chao Shen, Yung-Yu Chuang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3057-3066

Recent multi-view stereo methods have achieved promising results with the advancement of deep learning techniques. Despite of the progress, due to the limited fields of view of regular images, reconstructing large indoor environments still requires collecting many images with sufficient visual overlap, which is quite labor-intensive. 360deg images cover a much larger field of view than regular images and would facilitate the capture process. In this paper, we present 360MVSNet, the first deep learning network for multi-view stereo with 360deg images. Our method combines uncertainty estimation with a spherical sweeping module for 360deg images captured from multiple viewpoints in order to construct multi-scale cost volumes. By regressing volumes in a coarse-to-fine manner, high-resolution depth maps can be obtained. Furthermore, we have constructed EQMVS, a large-scale synthetic dataset that consists of over 50K pairs of RGB and depth maps in equirectangular projection. Experimental results demonstrate that our method can reconstruct large synthetic and real-world indoor scenes with significantly better completeness than previous traditional and learning-based methods while saving both time and effort in the data acquisition process.

**********

3D GAN Inversion With Pose Optimization

Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, Seungryong Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2967-2976

With the recent advances in NeRF-based 3D aware GANs quality, projecting an image into the latent space of these 3D-aware GANs has a natural advantage over 2D GAN inversion: not only does it allow multi-view consistent editing of the projected image, but it also enables 3D reconstruction and novel view synthesis when given only a single image. However, the explicit viewpoint control acts as a main hindrance in the 3D GAN inversion process, as both camera pose and latent code have to be optimized simultaneously to reconstruct the given image. Most works that explore the latent space of the 3D-aware GANs rely on ground-truth camera viewpoint or deformable 3D model, thus limiting their applicability. In this work, we introduce a generalizable 3D GAN inversion method that infers camera viewpoint and latent code simultaneously to enable multi-view consistent semantic image editing. The key to our approach is to leverage pre-trained estimators for better initialization and utilize the pixel-wise depth calculated from NeRF parameters to better reconstruct the given image. We conduct extensive experiments on image reconstruction and editing both quantitatively and qualitatively, and further compare our results with 2D GAN-based editing to demonstrate the advantages of utilizing the latent space of 3D GANs. Additional results and visualizations are available at https://hypernerf.github.io/.

**********

Improving Predicate Representation in Scene Graph Generation by Self-Supervised

Learning
So Hasegawa, Masayuki Hiromoto, Akira Nakagawa, Yuhei Umeda; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2740-2749
Scene graph generation (SGG) aims to understand sophisticated visual information by detecting triplets of subject, object, and their relationship (predicate). Since the predicate labels are heavily imbalanced, existing supervised methods struggle to improve accuracy for the rare predicates due to insufficient labeled data. In this paper, we propose SePiR, a novel self-supervised learning method for SGG to improve the representation of rare predicates. We first train a relational encoder by contrastive learning without using predicate labels, and then fine-tune a predicate classifier with labeled data. To apply contrastive learning to SGG, we newly propose data augmentation in which subject-object pairs are augmented by replacing their visual features with those from other images having the same object labels. By such augmentation, we can increase the variation of the visual features while keeping the relationship between the objects. Comprehensive experimental results on the Visual Genome dataset show that the SGG performance of SePiR is comparable to the state-of-the-art, and especially with the limited labeled dataset, our method significantly outperforms the existing supervised methods. Moreover, SePiR's improved representation enables the model architecture simpler, resulting in 3.6x and 6.3x reduction of the parameters and inference time from the existing method, independently.
**********
Spike-Based Anytime Perception
Matthew Dutson, Yin Li, Mohit Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5294-5304
In many emerging computer vision applications, it is critical to adhere to stringent latency and power constraints. The current neural network paradigm of frame-based, floating-point inference is often ill-suited to these resource-constrained applications. Spike-based perception - enabled by spiking neural networks (SNNs) - is one promising alternative. Unlike conventional neural networks (ANNs), spiking networks exhibit smooth tradeoffs between latency, power, and accuracy. SNNs are the archetype of an "anytime algorithm" whose accuracy improves smoothly over time. This property allows SNNs to adapt their computational investment in response to changing resource constraints. Unfortunately, mainstream algorithms for training SNNs (i.e., those based on ANN-to-SNN conversion) tend to produce models that are inefficient in practice. To mitigate this problem, we propose a set of principled optimizations that reduce latency and power consumption by 1-2 orders of magnitude in converted SNNs. These optimizations leverage a set of novel efficiency metrics designed for anytime algorithms. We also develop a state-of-the-art simulator, SaRNN, which can simulate SNNs using commodity GPU hardware and neuromorphic platforms. We hope that the proposed optimizations, metrics, and tools will facilitate the future development of spike-based vision systems.
**********
My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization
Umur A. Çiftçi, Gokturk Yuksek, ■lke Demir; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1369-1379
Recently, productization of face recognition and identification algorithms have become the most controversial topic about ethical AI. As new policies around digital identities are formed, we introduce three face access models in a hypothetical social network, where the user has the power to only appear in photos they approve. Our approach eclipses current tagging systems and replaces unapproved faces with quantitatively dissimilar deepfakes. In addition, we propose new metrics specific for this task, where the deepfake is generated at random with a guaranteed dissimilarity. We explain access models based on strictness of the data flow, and discuss impact of each model on privacy, usability, and performance. We evaluate our system on Facial Descriptor Dataset as the real dataset, and two synthetic datasets with random and equal class distributions. Running seven SOTA face recognizers on our results, MFMC reduces the average accuracy by 61%. Lastly, we extensively analyze similarity metrics, deepfake generators, and datasets i

n structural, visual, and generative spaces; supporting the design choices and v
erifying the quality.
**********

Difficulty-Net: Learning To Predict Difficulty for Long-Tailed Recognition
Saptarshi Sinha, Hiroki Ohashi; Proceedings of the IEEE/CVF Winter Conference on
 Applications of Computer Vision (WACV), 2023, pp. 6444-6453
Long-tailed datasets, where head classes comprise much more training samples tha
n tail classes, cause recognition models to get biased towards the head classes.
 Weighted loss is one of the most popular ways of mitigating this issue, and a r
ecent work has suggested that class-difficulty might be a better clue than conve
ntionally used class-frequency to decide the distribution of weights. A heuristi
c formulation was used in the previous work for quantifying the difficulty, but
we empirically find that the optimal formulation varies depending on the charact
eristics of datasets. Therefore, we propose Difficulty-Net, which learns to pred
ict the difficulty of classes using the model's performance in a meta-learning f
ramework. To make it learn reasonable difficulty of a class within the context o
f other classes, we newly introduce two key concepts, namely the relative diffic
ulty and the driver loss. The former helps Difficulty-Net take other classes int
o account when calculating difficulty of a class, while the latter is indispensa
ble for guiding the learning to a meaningful direction. Extensive experiments on
 popular long-tailed datasets demonstrated the effectiveness of the proposed met
hod, and it achieved state-of-the-art performance on multiple long-tailed datase
ts. Code is available at https://github.com/hitachi-rd-cv/Difficulty_Net.
**********

Scaling Neural Face Synthesis to High FPS and Low Latency by Neural Caching
Frank Yu, Sid Fels, Helge Rhodin; Proceedings of the IEEE/CVF Winter Conference
on Applications of Computer Vision (WACV), 2023, pp. 3474-3483
Recent neural rendering approaches greatly improve image quality, reaching near
photorealism. However, the underlying neural networks have high runtime, preclud
ing telepresence and virtual reality applications that require high resolution a
t low latency. The sequential dependency of layers in deep networks makes their
optimization difficult. We break this dependency by caching information from the
 previous frame to speed up the processing of the current one with an implicit w
arp. The warping with a shallow network reduces latency and the caching operatio
ns can further be parallelized to improve the frame rate. In contrast to existin
g temporal neural networks, ours is tailored for the task of rendering novel vie
ws of faces by conditioning on the change of the underlying surface mesh. We tes
t the approach on view-dependent rendering of 3D portrait avatars, as needed for
 telepresence, on established benchmark sequences. Warping reduces latency by 70
% (from 49.4ms to 14.9ms on commodity GPUs) and scales frame rates accordingly o
ver multiple GPUs while reducing image quality by only 1%, making it suitable as
 part of end-to-end view-dependent 3D teleconferencing applications.
**********

Single Image Super-Resolution via a Dual Interactive Implicit Neural Network
Quan H. Nguyen, William J. Beksi; Proceedings of the IEEE/CVF Winter Conference
on Applications of Computer Vision (WACV), 2023, pp. 4936-4945
In this paper, we introduce a novel implicit neural network for the task of sing
le image super-resolution at arbitrary scale factors. To do this, we represent a
n image as a decoding function that maps locations in the image along with their
 associated features to their reciprocal pixel attributes. Since the pixel locat
ions are continuous in this representation, our method can refer to any location
 in an image of varying resolution. To retrieve an image of a particular resolut
ion, we apply a decoding function to a grid of locations each of which refers to
 the center of a pixel in the output image. In contrast to other techniques, our
 dual interactive neural network decouples content and positional features. As a
 result, we obtain a fully implicit representation of the image that solves the
super-resolution problem at (real-valued) elective scales using a single model.
We demonstrate the efficacy and flexibility of our approach against the state of
 the art on publicly available benchmark datasets.
**********

Probabilistic Integration of Object Level Annotations in Chest X-Ray Classification
Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3630-3640
Medical image datasets and their annotations are not growing as fast as their equivalents in the general domain. This makes translation from the newest, more data-intensive methods that have made a large impact on the vision field increasingly more difficult and less efficient. In this paper, we propose a new probabilistic latent variable model for disease classification in chest X-ray images. Specifically we consider chest X-ray datasets that contain global disease labels, and for a smaller subset contain object level expert annotations in the form of eye gaze patterns and disease bounding boxes. We propose a two-stage optimization algorithm which is able to handle these different label granularities through a single training pipeline in a two-stage manner. In our pipeline global dataset features are learned in the lower level layers of the model. The specific details and nuances in the fine-grained expert object-level annotations are learned in the final layers of the model using a knowledge distillation method inspired by conditional variational inference. Subsequently, model weights are frozen to guide this learning process and prevent overfitting on the smaller richly annotated data subsets. The proposed method yields consistent classification improvement across different backbones on the common benchmark datasets Chest X-ray14 and MIMIC-CXR. This shows how two-stage learning of labels from coarse to fine-grained, in particular with object level annotations, is an effective method for more optimal annotation usage.
**********
FLAVR: Flow-Agnostic Video Representations for Fast Frame Interpolation
Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, Du Tran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2071-2082
Most modern frame interpolation approaches rely on explicit bidirectional optical flows between adjacent frames, thus are sensitive to the accuracy of underlying flow estimation in handling occlusions while additionally introducing computational bottlenecks unsuitable for efficient deployment. In this work, we propose a flow-free approach that is completely end-to-end trainable for multi-frame video interpolation. Our method, FLAVR, is designed to reason about non-linear motion trajectories and complex occlusions implicitly from unlabeled videos and greatly simplifies the process of training, testing and deploying frame interpolation models. Furthermore, FLAVR delivers up to 6x speed up compared to the current state-of-the-art methods for multi-frame interpolation while consistently demonstrating superior qualitative and quantitative results compared with prior methods on popular benchmarks including Vimeo-90K, Adobe-240FPS, and GoPro. Finally, we show that frame interpolation is a competitive self-supervised pre-training task for videos via demonstrating various novel applications of FLAVR including action recognition, optical flow estimation, motion magnification, and video object tracking. Code and trained models will be publicly released.
**********
Li3DeTr: A LiDAR Based 3D Detection Transformer
Gopi Krishna Erabati, Helder Araujo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4250-4259
Inspired by recent advances in vision transformers for object detection, we propose Li3DeTr, an end-to-end LiDAR based 3D Detection Transformer for autonomous driving, that inputs LiDAR point clouds and regresses 3D bounding boxes. The LiDAR local and global features are encoded using sparse convolution and multi-scale deformable attention respectively. In the decoder head, firstly, in the novel Li3DeTr cross-attention block, we link the LiDAR global features to 3D predictions leveraging the sparse set of object queries learnt from the data. Secondly, the object query interactions are formulated using multi-head self-attention. Finally, the decoder layer is repeated Ldec number of times to refine the object queries. Inspired by DETR, we employ set-to-set loss to train the Li3DeTr network.

Without bells and whistles, the Li3DeTr network achieves 61.3% mAP and 67.6% NDS surpassing the state-of-the-art methods with non-maximum suppression (NMS) on the nuScenes dataset and it also achieves competitive performance on the KITTI dataset. We also employ knowledge distillation (KD) using a teacher and student model that slightly improves the performance of our network.
**********

Efficient Visual Tracking With Exemplar Transformers
Philippe Blatter, Menelaos Kanakis, Martin Danelljan, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1571-1581
The design of more complex and powerful neural network models has significantly advanced the state-of-the-art in visual object tracking. These advances can be attributed to deeper networks, or the introduction of new building blocks, such as transformers. However, in the pursuit of increased tracking performance, runtime is often hindered. Furthermore, efficient tracking architectures have received surprisingly little attention. In this paper, we introduce the Exemplar Transformer, a transformer module utilizing a single instance level attention layer for realtime visual object tracking. E.T.Track, our visual tracker that incorporates Exemplar Transformer modules, runs at 47 FPS on a CPU. This is up to 8x faster than other transformer-based models. When compared to lightweight trackers that can operate in realtime on standard CPUs, E.T.Track consistently outperforms all other methods on the LaSOT, OTB-100, NFS, TrackingNet, and VOT-ST2020 datasets. Code and models are available at https://github.com/pblatter/ettrack.
**********

Kinematic-Aware Hierarchical Attention Network for Human Pose Estimation in Videos
Kyung-Min Jin, Byoung-Sung Lim, Gun-Hee Lee, Tae-Kyung Kang, Seong-Whan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5725-5734
Previous video-based human pose estimation methods have shown promising results by leveraging aggregated features of consecutive frames. However, most approaches compromise accuracy to mitigate jitter or do not sufficiently comprehend the temporal aspects of human motion. Furthermore, occlusion increases uncertainty between consecutive frames, which results in unsmooth results. To address these issues, we design an architecture that exploits the keypoint kinematic features with the following components. First, we effectively capture the temporal features by leveraging individual keypoint's velocity and acceleration. Second, the proposed hierarchical transformer encoder aggregates spatio-temporal dependencies and refines the 2D or 3D input pose estimated from existing estimators. Finally, we provide an online cross-supervision between the refined input pose generated from the encoder and the final pose from our decoder to enable joint optimization. We demonstrate comprehensive results and validate the effectiveness of our model in various tasks: 2D pose estimation, 3D pose estimation, body mesh recovery, and sparsely annotated multi-human pose estimation. Our code is available at https://github.com/KyungMinJin/HANet.
**********

Audio-Visual Face Reenactment
Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5178-5187
This work proposes a novel method to generate realistic talking head videos using audio and visual streams. We animate a source image by transferring head motion from a driving video using a dense motion field generated using learnable keypoints. We improve the quality of lip sync using audio as an additional input, helping the network to attend to the mouth region. We use additional priors using face segmentation and face mesh to improve the structure of the reconstructed faces. Finally, we improve the visual quality of the generations by incorporating a carefully designed identity-aware generator module. The identity-aware generator takes the source image and the warped motion features as input to generate a high-quality output with fine-grained details. Our method produces state-of-the-

art results and generalizes well to unseen faces, languages, and voices. We comprehensively evaluate our approach using multiple metrics and outperforming the current techniques both qualitative and quantitatively. Our work opens up several applications, including enabling low bandwidth video calls. We release a demo video and additional information at http://cvit.iiit.ac.in/research/projects/cvit-projects/avfr

**********

SALAD: Source-Free Active Label-Agnostic Domain Adaptation for Classification, Segmentation and Detection

Divya Kothandaraman, Sumit Shekhar, Abhilasha Sancheti, Manoj Ghuhan, Tripti Shukla, Dinesh Manocha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 382-391

We present a novel method, SALAD, for the challenging vision task of adapting a pre-trained "source" domain network to a "target" domain, with a small budget for annotation in the "target" domain and a shift in the label space. Further, the task assumes that the source data is not available for adaptation, due to privacy concerns or otherwise. We postulate that such systems need to jointly optimize the dual task of (i) selecting fixed number of samples from the target domain for annotation and (ii) transfer of knowledge from the pre-trained network to the target domain. To do this, SALAD consists of a novel Guided Attention Transfer Network (GATN) and an active learning function, HAL. The GATN enables feature distillation from pre-trained network to the target network, complemented with the target samples mined by HAL using transfer-ability and uncertainty criteria. SALAD has three key benefits: (i) it is task-agnostic, and can be applied across various visual tasks such as classification, segmentation and detection; (ii) it can handle shifts in output label space from the pre-trained source network to the target domain; (iii) it does not require access to source data for adaptation. We conduct extensive experiments across 3 visual tasks, viz. digits classification (MNIST, SVHN, VISDA), synthetic (GTA5) to real (CityScapes) image segmentation, and document layout detection (PubLayNet to DSSE). We show that our source-free approach, SALAD, results in an improvement of 0.5%-31.3% (across datasets and tasks) over prior adaptation methods that assume access to large amounts of annotated source data for adaptation.

**********

Large-Scale Open-Set Classification Protocols for ImageNet

Andres Palechor, Annesha Bhoumik, Manuel Günther; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 42-51

Open-Set Classification (OSC) intends to adapt closed-set classification models to real-world scenarios, where the classifier must correctly label samples of known classes while rejecting previously unseen unknown samples. Only recently, research started to investigate on algorithms that are able to handle these unknown samples correctly. Some of these approaches address OSC by including into the training set negative samples that a classifier learns to reject, expecting that these data increase the robustness of the classifier on unknown classes. Most of these approaches are evaluated on small-scale and low-resolution image datasets like MNIST, SVHN or CIFAR, which makes it difficult to assess their applicability to the real world, and to compare them among each other. We propose three open-set protocols that provide rich datasets of natural images with different levels of similarity between known and unknown classes. The protocols consist of subsets of ImageNet classes selected to provide training and testing data closer to real-world scenarios. Additionally, we propose a new validation metric that can be employed to assess whether the training of deep learning models addresses both the classification of known samples and the rejection of unknown samples. We use the protocols to compare the performance of two baseline open-set algorithms to the standard SoftMax baseline and find that the algorithms work well on negative samples that have been seen during training, and partially on out-of-distribution detection tasks, but drop performance in the presence of samples from previously unseen unknown classes.

**********

Probabilistic Volumetric Fusion for Dense Monocular SLAM

Antoni Rosinol, John J. Leonard, Luca Carlone; Proceedings of the IEEE/CVF Winte
r Conference on Applications of Computer Vision (WACV), 2023, pp. 3097-3105
We present a novel method to reconstruct 3D scenes from images by leveraging dee
p dense monocular SLAM and fast uncertainty propagation. The proposed approach i
s able to 3D reconstruct scenes densely, accurately, and in real-time while bein
g robust to extremely noisy depth estimates coming from dense monocular SLAM. Di
fferently from previous approaches, that either use ad-hoc depth filters, or tha
t estimate the depth uncertainty from RGB-D cameras' sensor models, our probabil
istic depth uncertainty derives directly from the information matrix of the unde
rlying bundle adjustment problem in SLAM. We show that the resulting depth uncer
tainty provides an excellent signal to weight the depth-maps for volumetric fusi
on. Without our depth uncertainty, the resulting mesh is noisy and with artifact
s, while our approach generates an accurate 3D mesh with significantly fewer art
ifacts. We provide results on the challenging Euroc dataset, and show that our a
pproach achieves 92% better accuracy than directly fusing depths from monocular
SLAM, and up to 90% improvements compared to the best competing approach.
**********

Learning 3D Human Pose Estimation From Dozens of Datasets Using a Geometry-Aware
 Autoencoder To Bridge Between Skeleton Formats
István Sárándi, Alexander Hermans, Bastian Leibe; Proceedings of the IEEE/CVF Wi
nter Conference on Applications of Computer Vision (WACV), 2023, pp. 2956-2966
Deep learning-based 3D human pose estimation performs best when trained on large
 amounts of labeled data, making combined learning from many datasets an importa
nt research direction. One obstacle to this endeavor are the different skeleton
formats provided by different datasets, i.e., they do not label the same set of
anatomical landmarks. There is little prior research on how to best supervise on
e model with such discrepant labels. We show that simply using separate output h
eads for different skeletons results in inconsistent depth estimates and insuffi
cient information sharing across skeletons. As a remedy, we propose a novel affi
ne-combining autoencoder (ACAE) method to perform dimensionality reduction on th
e number of landmarks. The discovered latent 3D points capture the redundancy am
ong skeletons, enabling enhanced information sharing when used for consistency r
egularization. Our approach scales to an extreme multi-dataset regime, where we
use 28 3D human pose datasets to supervise one model, which outperforms prior wo
rk on a range of benchmarks, including the challenging 3D Poses in the Wild (3DP
W) dataset. Our code and models are available for research purposes.
**********

Learning Classifiers of Prototypes and Reciprocal Points for Universal Domain Ad
aptation
Sungsu Hur, Inkyu Shin, Kwanyong Park, Sanghyun Woo, In So Kweon; Proceedings of
 the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023,
 pp. 531-540
Universal Domain Adaptation aims to transfer the knowledge between the datasets
by handling two shifts: domain-shift and category-shift. The main challenge is c
orrectly distinguishing the unknown target samples while adapting the distributi
on of known class knowledge from source to target. Most existing methods approac
h this problem by first training the target adapted known classifier and then re
lying on the single threshold to distinguish unknown target samples. However, th
is simple threshold-based approach prevents the model from considering the under
lying complexities existing between the known and unknown samples in the high-di
mensional feature space. In this paper, we propose a new approach in which we us
e two sets of feature points, namely dual Classifiers for Prototypes and Recipro
cals (CPR). Our key idea is to associate each prototype with corresponding known
 class features while pushing the reciprocals apart from these prototypes to loc
ate them in the potential unknown feature space. The target samples are then cla
ssified as unknown if they fall near any reciprocals at test time. To successful
ly train our framework, we collect the partial, confident target samples that ar
e classified as known or unknown through on our proposed multi-criteria selectio
n. We then additionally apply the entropy loss regularization to them. For furth
er adaptation, we also apply standard consistency regularization that matches th

e predictions of two different views of the input to make more compact target fe ature space. We evaluate our proposal, CPR, on three standard benchmarks and ach ieve comparable or new state-of-the-art results. We also provide extensive ablat ion experiments to verify our main design choices in our framework.
**********

GEMS: Generating Efficient Meta-Subnets
Varad Pimpalkhute, Shruti Kunde, Rekha Singhal; Proceedings of the IEEE/CVF Wint er Conference on Applications of Computer Vision (WACV), 2023, pp. 5315-5323
Gradient-based meta learners (GBML) such as MAML aim to learn a model initializa tion across similar tasks, such that the model generalizes well on unseen tasks sampled from the same distribution with few gradient updates. A limitation of GB ML is its inability to adapt to real-world applications where input tasks are sa mpled from multiple distributions. An existing effort learns N initializations f or tasks sampled from N distributions; roughly increasing training time by a fac tor of N. Instead, we use a single model initialization to learn distribution-sp ecific parameters for every input task. This reduces negative knowledge transfer across distributions and overall computational cost. Specifically, we explore t wo ways of efficiently learning on multi-distribution tasks: 1) Binary Mask Perc eptron (BMP), which learns distribution-specific layers, 2) Multi-modal Supermas k (MMSUP), which learns distribution-specific parameters. We evaluate the perfor mance of the proposed framework (GEMS) on few-shot vision classification tasks. The experimental results demonstrate a significant improvement in accuracy and r eduction in training time over existing state of the art algorithms on quasi-ben chmark tasks.
**********

Self-Supervised Distilled Learning for Multi-Modal Misinformation Identification
Michael Mu, Sreyasee Das Bhattacharjee, Junsong Yuan; Proceedings of the IEEE/CV F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2819-28 28
Rapid dissemination of misinformation is a major societal problem receiving incr easing attention. Unlike Deepfake, Out-of-Context misinformation, in which the u naltered unimode contents (e.g. image, text) of a multi-modal news sample are co mbined in an out-of-context manner to generate deception, requires limited techn ical expertise to create. Therefore, it is more prevalent a means to confuse rea ders. Most existing approaches extract features from its uni-mode counterparts t o concatenate and train a model for the misinformation classification task. In t his paper, we design a self-supervised feature representation learning strategy that aims to attain the multi-task objectives: (1) task-agnostic, which evaluate s the intra- and inter-mode representational consistencies for improved alignmen ts across related models; (2) task-specific, which estimates the category-specif ic multi-modal knowledge to enable the classifier to derive more discriminative predictive distributions. To compensate for the dearth of annotated data represe nting varied types of misinformation, the proposed Self-Supervised Distilled Lea rner (SSDL) utilizes a Teacher network to weakly guide a Student network to mimi c a similar decision pattern as the teacher. The two-phased learning of SSDL can be summarized as: initial pretraining of the Student model using a combination of contrastive self-supervised task-agnostic objective and supervised task-speci fic adjustment in parallel; finetuning the Student model via self-supervised kno wledge distillation blended with the supervised objective of decision alignment. In addition to the consistent out-performances over the existing baselines that demonstrate the feasibility of our approach, the explainability capacity of the proposed SSDL also helps users visualize the reasoning behind a specific predic tion made by the model.
**********

Tracking Growth and Decay of Plant Roots in Minirhizotron Images
Alexander Gillert, Bo Peters, Uwe Freiherr von Lukas, Jürgen Kreyling, Gesche Bl ume-Werry; Proceedings of the IEEE/CVF Winter Conference on Applications of Comp uter Vision (WACV), 2023, pp. 3699-3708
Plant roots are difficult to monitor and study since they are hidden belowground . Minirhizotrons offer an in-situ monitoring solution but their widespread adopt

ion is still limited by the capabilities of automatic analysis methods. These ca
pabilities so far consist only of estimating a single number (total root length)
 per image.\nWe propose a method for a more fine-grained analysis which estimate
s the root turnover, i.e. the amount of root growth and decay between two minirh
izotron images. It consists of a neural network that computes which roots are vi
sible in both images and is trained in an unsupervised manner without additional
 annotations.\nOur code is available as a part of an analysis tool with a user i
nterface ready to be used by ecologists.
**********

LiveSeg: Unsupervised Multimodal Temporal Segmentation of Long Livestream Videos
Jielin Qiu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Ding Zhao, Hailin Jin;
Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision
 (WACV), 2023, pp. 5188-5198
Livestream videos have become a significant part of online learning, where desig
n, digital marketing, creative painting, and other skills are taught by experien
ced experts in the sessions, making them valuable materials. However, Livestream
 tutorial videos are usually hours long, recorded, and uploaded to the Internet
directly after the live sessions, making it hard for other people to catch up qu
ickly. An outline will be a beneficial solution, which requires the video to be
temporally segmented according to topics. In this work, we introduced a large Li
vestream video dataset named MultiLive, and formulated the temporal segmentation
 of the long Livestream videos (TSLLV) task. We propose LiveSeg, an unsupervised
 Livestream video temporal Segmentation solution, which takes advantage of multi
modal features from different domains. Our method achieved a 16.8% F1-score perf
ormance improvement compared with the state-of-the-art method.
**********

Enhanced Bi-Directional Motion Estimation for Video Frame Interpolation
Xin Jin, Longhai Wu, Guotao Shen, Youxin Chen, Jie Chen, Jayoon Koo, Cheul-hee H
ahm; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer V
ision (WACV), 2023, pp. 5049-5057
We propose a simple yet effective algorithm for motion-based video frame interpo
lation. Existing motion-based interpolation methods typically rely on an off-the
-shelf optical flow model or a U-Net based pyramid network for motion estimation
, which either suffer from large model size or limited capacity in handling vari
ous challenging motion cases. In this work, we present a novel compact model to
simultaneously estimate the bi-directional motions between input frames. It is d
esigned by carefully adapting the ingredients (e.g., warping, correlation) in op
tical flow research for simultaneous bi-directional motion estimation within a f
lexible pyramid recurrent framework. Our motion estimator is extremely lightweig
ht (15x smaller than PWC-Net), yet enables reliable handling of large and comple
x motion cases. Based on estimated bi-directional motions, we employ a synthesis
 network to fuse forward-warped representations and predict the intermediate fra
me. Our method achieves excellent performance on a broad range of frame interpol
ation benchmarks. Code and trained models are available at: https://github.com/s
rcn-ivl/EBME.
**********

LoopDA: Constructing Self-Loops To Adapt Nighttime Semantic Segmentation
Fengyi Shen, Zador Pataki, Akhil Gurram, Ziyuan Liu, He Wang, Alois Knoll; Proce
edings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC
V), 2023, pp. 3256-3266
Due to the lack of training labels and the difficulty of annotating, dealing wit
h adverse driving conditions such as nighttime has posed a huge challenge to the
 perception system of autonomous vehicles. Therefore, adapting knowledge from a
labelled daytime domain to an unlabelled nighttime domain has been widely resear
ched. In addition to labelled daytime datasets, existing nighttime datasets usua
lly provide nighttime images with corresponding daytime reference images capture
d at nearby locations for reference. The key challenge is to minimize the perfor
mance gap between the two domains. In this paper, we propose LoopDA for domain a
daptive nighttime semantic segmentation. It consists of self-loops that result i
n reconstructing the input data using predicted semantic maps, by rendering them

into the encoded features. In a warm-up training stage, the self-loops comprise of an inner-loop and an outer-loop, which are responsible for intra-domain refinement and inter-domain alignment, respectively. To reduce the impact of day-night pose shifts, in the later self-training stage, we propose a co-teaching pipeline that involves an offline pseudo-supervision signal and an online reference-guided signal 'DNA' (Day-Night Agreement), bringing substantial benefits to enhance nighttime segmentation. Our model outperforms prior methods on Dark Zurich and Nighttime Driving datasets for semantic segmentation. Code and pretrained models are available at https://github.com/fy-vision/LoopDA.
**********

Efficient Reference-Based Video Super-Resolution (ERVSR): Single Reference Image Is All You Need
Youngrae Kim, Jinsu Lim, Hoonhee Cho, Minji Lee, Dongman Lee, Kuk-Jin Yoon, Ho-Jin Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1828-1837
Reference-based video super-resolution (RefVSR) is a promising domain of super-resolution that recovers high-frequency textures of a video using reference video. The multiple cameras with different focal lengths in mobile devices aid recent works in RefVSR, which aim to super-resolve a low-resolution ultra-wide video by utilizing wide-angle videos. Previous works in RefVSR used all reference frames of a Ref video at each time step for the super-resolution of low-resolution videos. However, computation on higher-resolution images increases the runtime and memory consumption, hence hinders the practical application of RefVSR. To solve this problem, we propose an Efficient Reference-based Video Super-Resolution (ERVSR) that exploits a single reference frame to super-resolve whole low-resolution video frames. We introduce an attention-based feature align module and an aggregation upsampling module that attends LR features using the correlation between the reference and LR frames. The proposed ERVSR achieves 12xfaster speed, 1/4 memory consumption than previous state-of-the-art RefVSR networks, and competitive performance on the RealMCVSR dataset while using a single reference image.
**********

Exploiting Visual Context Semantics for Sound Source Localization
Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, Wanli Ouyang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5199-5208
Self-supervised sound source localization in unconstrained visual scenes is an important task of audio-visual learning. In this paper, we propose a visual reasoning module to explicitly exploit the rich visual context semantics, which alleviates the issue of insufficient utilization of visual information in previous works. The learning objectives are carefully designed to provide stronger supervision signals for the extracted visual semantics while enhancing the audio-visual interactions, which lead to more robust feature representations. Extensive experimental results demonstrate that our approach significantly boosts the localization performances on various datasets, even without initializations pretrained on ImageNet. Moreover, with the visual context exploitation, our framework can accomplish both the audio-visual and purely visual inference, which expands the application scope of the sound source localization task and further raises the competitiveness of our approach.
**********

Improving the Robustness of Point Convolution on K-Nearest Neighbor Neighborhoods With a Viewpoint-Invariant Coordinate Transform
Xingyi Li, Wenxuan Wu, Xiaoli Z. Fern, Li Fuxin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1287-1297
Recently, there is significant interest in performing convolution over irregularly sampled point clouds. Point clouds are very different from raster images, in that one cannot have a regular sampling grid on point clouds, which makes robustness under irregular neighborhoods an important issue. Especially, the k-nearest neighbor (kNN) neighborhood presents challenges for generalization because the location of the neighbors can be very different between training and testing times. In order to improve the robustness to different neighborhood samplings, this

paper proposes a novel viewpoint-invariant coordinate transform as the input to the weight-generating function for point convolution, in addition to the regular 3D coordinates. This allows us to feed the network with non-invariant, scale-invariant and scale+rotation-invariant coordinates simultaneously, so that the network can learn which to include in the convolution function automatically. Empirically, we demonstrate that this effectively improves the performance of point cloud convolutions on the SemanticKITTI and ScanNet datasets, as well as the robustness to significant test-time downsampling, which can substantially change the distance of neighbors in a kNN neighborhood. Experimentally, among pure point-based approaches, we achieve comparable semantic segmentation performance with a comparable point-based convolution framework KPConv on SemanticKITTI and ScanNet, yet is significantly more efficient by virtue of using a kNN neighborhood instead of an -ball.

**********

Robust and Efficient Alignment of Calcium Imaging Data Through Simultaneous Low Rank and Sparse Decomposition

Junmo Cho, Seungjae Han, Eun-Seo Cho, Kijung Shin, Young-Gyu Yoon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1939-1948

Accurate alignment of calcium imaging data, which is critical for the extraction of neuronal activity signals, is often hindered by the image noise and the neuronal activity itself. To address the problem, we propose an algorithm named REALS for robust and efficient batch image alignment through simultaneous geometric transformation and low rank and sparse decomposition. REALS is constructed upon our finding that the low rank subspace can be recovered via linear projection, which allows us to perform simultaneous image alignment and decomposition with gradient-based updates. REALS achieves orders-of magnitude improvement in terms of accuracy and speed compared to the state-of-the-art robust image alignment algorithms. In addition, we introduce two extended versions of REALS that achieve even higher accuracy than REALS under challenging conditions. First, multi-resolution REALS achieves up to 5 times higher alignment accuracy than REALS. Second, deformable REALS generalizes REALS for non-rigid registration. Furthermore, REALS can be combined with downstream tasks such as unsupervised image segmentation owing to its differentiability.

**********

FreeREA: Training-Free Evolution-Based Architecture Search

Niccolò Cavagnero, Luca Robbiano, Barbara Caputo, Giuseppe Averta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1493-1502

In the last decade, most research in Machine Learning contributed to the improvement of existing models, with the aim of increasing the performance of neural networks for the solution of a variety of different tasks. However, such advancements often come at the cost of an increase of model memory and computational requirements. This represents a significant limitation for the deployability of research output in realistic settings, where the cost, the energy consumption, and the complexity of the framework play a crucial role. To solve this issue, the designer should search for models that maximise the performance while limiting its footprint. Typical approaches to reach this goal rely either on manual procedures, which cannot guarantee the optimality of the final design, or upon Neural Architecture Search algorithms to automatise the process, at the expenses of extremely high computational time. This paper provides a solution for the fast identification of a neural network that maximises the model accuracy while preserving size and computational constraints typical of tiny devices. Our approach, named FreeREA, is a custom cell-based evolution NAS algorithm that exploits an optimised combination of training-free metrics to rank architectures during the search, thus without need of model training. Our experiments, carried out on the common benchmarks NAS-Bench-101 and NATS-Bench, demonstrate that i) FreeREA is a fast, efficient and effective search method for models automatic design; ii) it outperforms State of the Art training-based and training-free techniques in all the datasets and benchmarks considered, and iii) it can easily generalise to constrain

ed scenarios, representing a competitive solution for fast Neural Architecture Search in generic constrained applications. The code is available at https://github.com/NiccoloCavagnero/FreeREA.
**********

Explainability-Aware One Point Attack for Point Cloud Neural Networks
Hanxiao Tan, Helena Kotthaus; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4581-4590
Recent studies have shown an increased interest to investigate the reliability of point cloud networks by adversarial attacks. However, most of the existing studies aim to deceive humans, while few address the operation principles of the models themselves. In this work, we propose two adversarial methods: One Point Attack (OPA) and Critical Traversal Attack (CTA), which target the points crucial to predictions more precisely by incorporating explainability methods. Our results show that popular point cloud networks can be deceived with almost 100% success rate by shifting only one point from the input instance. We also show the interesting impact of different point attribution distributions on the adversarial robustness of point cloud networks. We discuss how our approaches facilitate the explainability study for point cloud networks. To the best of our knowledge, this is the first point-cloud-based adversarial approach concerning explainability. Our code is available at https://github.com/Explain3D/Exp-One-Point-Atk-PC.
**********

Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion
Shruti Agarwal, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, Anna Rohrbach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4710-4719
In today's era of digital misinformation, we are increasingly faced with new threats posed by video falsification techniques. Such falsifications range from cheapfakes (e.g., lookalikes or audio dubbing) to deepfakes (e.g., sophisticated AI media synthesis methods), which are becoming perceptually indistinguishable from real videos. To tackle this challenge, we propose a multi-modal semantic forensic approach to discover clues that go beyond detecting discrepancies in visual quality, thereby handling both simpler cheapfakes and visually persuasive deepfakes. In this work, our goal is to verify that the purported person seen in the video is indeed themselves by detecting anomalous facial movements corresponding to the spoken words. We leverage the idea of attribution to learn person-specific biometric patterns that distinguish a given speaker from others. We use interpretable Action Units (AUs) to capture a persons' face and head movement as opposed to deep CNN features, and we are the first to use word-conditioned facial motion analysis. We further demonstrate our method's effectiveness on a range of fakes not seen in training including those without video manipulation, that were not addressed in prior work.
**********

Learnable Human Mesh Triangulation for 3D Human Pose and Shape Estimation
Sungho Chun, Sungbum Park, Ju Yong Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2850-2859
Compared to joint position, the accuracy of joint rotation and shape estimation has received relatively little attention in the skinned multi-person linear model (SMPL)-based human mesh reconstruction from multi-view images. The work in this field is broadly classified into two categories. The first approach performs joint estimation and then produces SMPL parameters by fitting SMPL to resultant joints. The second approach regresses SMPL parameters directly from the input images through a convolutional neural network (CNN)-based model. However, these approaches suffer from the lack of information for resolving the ambiguity of joint rotation and shape reconstruction and the difficulty of network learning. To solve the aforementioned problems, we propose a two-stage method. The proposed method first estimates the coordinates of mesh vertices through a CNN-based model from input images, and acquires SMPL parameters by fitting the SMPL model to the estimated vertices. Estimated mesh vertices provide sufficient information for determining joint rotation and shape, and are easier to learn than SMPL parameter

s. According to experiments using Human3.6M and MPI-INF-3DHP datasets, the propo sed method significantly outperforms the previous works in terms of joint rotati on and shape estimation, and achieves competitive performance in terms of joint location estimation.
**********

Multi-View Tracking Using Weakly Supervised Human Motion Prediction
Martin Engilberge, Weizhe Liu, Pascal Fua; Proceedings of the IEEE/CVF Winter Co nference on Applications of Computer Vision (WACV), 2023, pp. 1582-1592
Multi-view approaches to people-tracking have the potential to better handle occ lusions than single-view ones in crowded scenes. They often rely on the tracking -by-detection paradigm, which involves detecting people first and then connectin g the detections. In this paper, we argue that an even more effective approach i s to predict people motion over time and infer people's presence in individual f rames from these. This enables to enforce consistency both over time and across views of a single temporal frame. We validate our approach on the PETS2009 and W ILDTRACK datasets and demonstrate that it outperforms state-of-the-art methods.
**********

Augmentation by Counterfactual Explanation - Fixing an Overconfident Classifier
Sumedha Singla, Nihal Murali, Forough Arabshahi, Sofia Triantafyllou, Kayhan Bat manghelich; Proceedings of the IEEE/CVF Winter Conference on Applications of Com puter Vision (WACV), 2023, pp. 4720-4730
A highly accurate but overconfident model is ill-suited for deployment in critic al applications such as healthcare and autonomous driving. The classification ou tcome should reflect a high uncertainty on ambiguous in-distribution samples tha t lie close to the decision boundary. The model should also refrain from making overconfident decisions on samples that lie far outside its training distributio n, far-out-of-distribution (far-OOD), or on unseen samples from novel classes th at lie near its training distribution (near-OOD). This paper proposes an applica tion of counterfactual explanations in fixing an over-confident classifier. Spec ifically, we propose to fine-tune a given pre-trained classifier using augmentat ions from a counterfactual explainer (ACE) to fix its uncertainty characteristic s while retaining its predictive performance. We perform extensive experiments w ith detecting far-OOD, near-OOD, and ambiguous samples. Our empirical results sh ow that the revised model have improved uncertainty measures, and its performanc e is competitive to the state-of-the-art methods.
**********

IFQA: Interpretable Face Quality Assessment
Byungho Jo, Donghyeon Cho, In Kyu Park, Sungeun Hong; Proceedings of the IEEE/CV F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3444-34 53
Existing face restoration models have relied on general assessment metrics that do not consider the characteristics of facial regions. Recent works have therefo re assessed their methods using human studies, which is not scalable and involve s significant effort. This paper proposes a novel face-centric metric based on a n adversarial framework where a generator simulates face restoration and a discr iminator assesses image quality. Specifically, our per-pixel discriminator enabl es interpretable evaluation that cannot be provided by traditional metrics. More over, our metric emphasizes facial primary regions considering that even minor c hanges to the eyes, nose, and mouth significantly affect human cognition. Our fa ce-oriented metric consistently surpasses existing general or facial image quali ty assessment metrics by impressive margins. We demonstrate the generalizability of the proposed strategy in various architectural designs and challenging scena rios. Interestingly, we find that our IFQA can lead to performance improvement a s an objective function. The code and models are available at https://github.com /VCLLab/IFQA.
**********

High-Quality RGB-D Reconstruction via Multi-View Uncalibrated Photometric Stereo and Gradient-SDF
Lu Sang, Björn Häfner, Xingxing Zuo, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3106-311

5
Fine-detailed reconstructions are in high demand in many applications. However, most of the existing RGB-D reconstruction methods rely on pre-calculated accurate camera poses to recover the detailed surface geometry, where the representation of a surface needs to be adapted when optimizing different quantities. In this paper, we present a novel multi-view RGB-D based reconstruction method that tackles camera pose, lighting, albedo, and surface normal estimation via the utilization of a gradient signed distance field (Gradient-SDF). The proposed method formulates the image rendering process using specific physically-based model(s) and optimizes the surface's quantities on the actual surface using its volumetric representation, as opposed to other works which estimate surface quantities only near the actual surface. To validate our method, we investigate two physically-based image formation models for natural light and point light source applications. The experimental results on synthetic and real-world datasets demonstrate that the proposed method can recover high-quality geometry of the surface more faithfully than the state-of-art and further improves the accuracy of estimated camera poses.
**********

NeuralBF: Neural Bilateral Filtering for Top-Down Instance Segmentation on Point Clouds
Weiwei Sun, Daniel Rebain, Renjie Liao, Vladimir Tankovich, Soroosh Yazdani, Kwang Moo Yi, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 551-560
We introduce a method for instance proposal generation for 3D point clouds. Existing techniques typically directly regress proposals in a single feed-forward step, leading to inaccurate estimation. We show that this serves as a critical bottleneck, and propose a method based on iterative bilateral filtering with learned kernels. Following the spirit of bilateral filtering, we consider both the deep feature embeddings of each point, as well as their locations in the 3D space. We show via synthetic experiments that our method brings drastic improvements when generating instance proposals for a given point of interest. We further validate our method on the challenging ScanNet benchmark, achieving the best instance segmentation performance amongst the sub-category of top-down methods.
**********

UVCGAN: UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation
Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, Yihui Ren; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 702-712
Unpaired image-to-image translation has broad applications in art, design, and scientific simulations. One early breakthrough was CycleGAN that emphasizes one-to-one mappings between two unpaired image domains via generative-adversarial networks (GAN) coupled with the cycle-consistency constraint, while more recent works promote one-to-many mapping to boost diversity of the translated images. Motivated by scientific simulation and one-to-one needs, this work revisits the classic CycleGAN framework and boosts its performance to outperform more contemporary models without relaxing the cycle-consistency constraint. To achieve this, we equip the generator with a Vision Transformer (ViT) and employ necessary training and regularization techniques. Compared to previous best-performing models, our model performs better and retains a strong correlation between the original and translated image. An accompanying ablation study shows that both the gradient penalty and self-supervised pre-training are crucial to the improvement. To promote reproducibility and open science, the source code, hyperparameter configurations, and pre-trained model are available at https: //github.com/LS4GAN/uvcgan.
**********

X-Align: Cross-Modal Cross-View Alignment for Bird's-Eye-View Segmentation
Shubhankar Borse, Marvin Klingner, Varun Ravi Kumar, Hong Cai, Abdulaziz Almuzairee, Senthil Yogamani, Fatih Porikli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3287-3297
Bird's-eye-view (BEV) grid is a common representation for the perception of road

components, e.g., drivable area, in autonomous driving. Most existing approaches rely on cameras only to perform segmentation in BEV space, which is fundamentally constrained by the absence of reliable depth information. Latest works leverage both camera and LiDAR modalities, but sub-optimally fuse their features using simple, concatenation-based mechanisms. In this paper, we address these problems by enhancing the alignment of the unimodal features in order to aid feature fusion, as well as enhancing the alignment between the cameras' perspective view (PV) and BEV representations. We propose X-Align, a novel end-to-end cross-modal and cross-view learning framework for BEV segmentation consisting of the following components: (i) a novel Cross-Modal Feature Alignment (X-FA) loss, (ii) an attention-based Cross-Modal Feature Fusion (X-FF) module to align multi-modal BEV features implicitly, and (iii) an auxiliary PV segmentation branch with Cross-View Segmentation Alignment (X-SA) losses to improve the PV-to-BEV transformation. We evaluate our proposed method across two commonly used benchmark datasets, i.e., nuScenes and KITTI-360. Notably, X-Align significantly outperforms the state-of-the-art by 3 absolute mIoU points on nuScenes. We also provide extensive ablation studies to demonstrate the effectiveness of the individual components.
**********

Domain Adaptive Object Detection for Autonomous Driving Under Foggy Weather
Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, Hongkai Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 612-622
Most object detection methods for autonomous driving usually assume a onsistent feature distribution between training and testing data, which is not always the case when weathers differ significantly. The object detection model trained under clear weather might be not effective enough on the foggy weather because of the domain gap. This paper proposes a novel domain adaptive object detection framework for autonomous driving under foggy weather. Our method leverages both image-level and object-level adaptation to diminish the domain discrepancy in image style and object appearance. To further enhance the model's capabilities under challenging samples, we also come up with a new adversarial gradient reversal layer to perform adversarial mining for the hard examples together with domain adaptation. Moreover, we propose to generate an auxiliary domain by data augmentation to enforce a new domain-level metric regularization. Experimental results on public benchmarks show the effectiveness and accuracy of the proposed method.
**********

Mobile Robot Manipulation Using Pure Object Detection
Brent Griffin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 561-571
This paper addresses the problem of mobile robot manipulation using object detection. Our approach uses detection and control as complimentary functions that learn from real-world interactions. We develop an end-to-end manipulation method based solely on detection and introduce Task-focused Few-shot Object Detection (TFOD) to learn new objects and settings. Our robot collects its own training data and automatically determines when to retrain detection to improve performance across various subtasks (e.g., grasping). Notably, detection training is low-cost, and our robot learns to manipulate new objects using as few as four clicks of annotation. In physical experiments, our robot learns visual control from a single click of annotation and a novel update formulation, manipulates new objects in clutter and other mobile settings, and achieves state-of-the-art results on an existing visual servo control and depth estimation benchmark. Finally, we develop a TFOD Benchmark to support future object detection research for robotics: https://github.com/griffbr/TFOD.
**********

FastSwap: A Lightweight One-Stage Framework for Real-Time Face Swapping
Sahng-Min Yoo, Tae-Min Choi, Jae-Woo Choi, Jong-Hwan Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3558-3567
Recent face swapping frameworks have achieved high-fidelity results. However, the previous works suffer from high computation costs due to the deep structure an

d the use of off-the-shelf networks. To overcome such problems and achieve real-time face swapping, we propose a lightweight one-stage framework, FastSwap. We design a shallow network trained in a self-supervised manner without any manual annotations. The core of our framework is a novel decoder block, called Triple Adaptive Normalization (TAN) block, which effectively integrates the identity and pose information. Besides, we propose a novel data augmentation and switch-test strategy to extract the attributes from the target image, which further enables controllable attribute editing. Extensive experiments on VoxCeleb2 and wild faces demonstrate that our framework generates high-fidelity face swapping results in 123.22 FPS and better preserves the identity, pose, and attributes than other state-of-the-art methods. Furthermore, we conduct an in-depth study to demonstrate the effectiveness of our proposal.

**********

Realistic Full-Body Anonymization With Surface-Guided GANs

Håkon Hukkelås, Morten Smebye, Rudolf Mester, Frank Lindseth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1430-1440

Recent work on image anonymization has shown that generative adversarial networks (GANs) can generate near-photorealistic faces to anonymize individuals. However, scaling up these networks to the entire human body has remained a challenging and yet unsolved task. We propose a new anonymization method that generates realistic humans for in-the-wild images. A key part of our design is to guide adversarial nets by dense pixel-to-surface correspondences between an image and a canonical 3D surface. We introduce Variational Surface-Adaptive Modulation (V-SAM) that embeds surface information throughout the generator. Combining this with our novel discriminator surface supervision loss, the generator can synthesize high quality humans with diverse appearances in complex and varying scenes. We demonstrate that surface guidance significantly improves image quality and diversity of samples, yielding a highly practical generator. Finally, we show that our method preserves data usability without infringing privacy when collecting image datasets for training computer vision models.

**********

ProtoSeg: Interpretable Semantic Segmentation With Prototypical Parts

Miko■aj Sacha, Dawid Rymarczyk, ■ukasz Struski, Jacek Tabor, Bartosz Zieli■ski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1481-1492

We introduce ProtoSeg, a novel model for interpretable semantic image segmentation, which constructs its predictions using similar patches from the training set. To achieve accuracy comparable to baseline methods, we adapt the mechanism of prototypical parts and introduce a diversity loss function that increases the variety of prototypes within each class. We show that ProtoSeg discovers semantic concepts, in contrast to standard segmentation models. Experiments conducted on Pascal VOC and Cityscapes datasets confirm the precision and transparency of the presented method.

**********

Structure-Encoding Auxiliary Tasks for Improved Visual Representation in Vision-and-Language Navigation

Chia-Wen Kuo, Chih-Yao Ma, Judy Hoffman, Zsolt Kira; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1104-1113

In Vision-and-Language Navigation (VLN), researchers typically take an image encoder pre-trained on ImageNet without fine-tuning on the environments that the agent will be trained or tested on. However, the distribution shift between the training images from ImageNet and the views in the navigation environments may render the ImageNet pre-trained image encoder suboptimal. Therefore, in this paper, we design a set of structure-encoding auxiliary tasks (SEA) that leverage the data in the navigation environments to pre-train and improve the image encoder. Specifically, we design and customize (1) 3D jigsaw, (2) traversability prediction, and (3) instance classification to pre-train the image encoder. Through rigorous ablations, our SEA pre-trained features are shown to better encode structura

l information of the scenes, which ImageNet pre-trained features fail to properl y encode but is crucial for the target navigation task. The SEA pre-trained feat ures can be easily plugged into existing VLN agents without any tuning. For exam ple, on Test-Unseen environments, the VLN agents combined with our SEA pre-train ed features achieve absolute success rate improvement of 12% for Speaker-Followe r [14], 5% for Env-Dropout [37], and 4% for AuxRN [50].
**********

Semantics-Depth-Symbiosis: Deeply Coupled Semi-Supervised Learning of Semantics and Depth
Nitin Bansal, Pan Ji, Junsong Yuan, Yi Xu; Proceedings of the IEEE/CVF Winter Co nference on Applications of Computer Vision (WACV), 2023, pp. 5828-5839
Multi-task learning (MTL) paradigm focuses on jointly learning two or more tasks , aiming for an improvement w.r.t model's generalizability, performance, and tra ining/inference memory footprint. The aforementioned benefits become ever so ind ispensable in the case of training for vision-related dense prediction tasks. In this work, we tackle the MTL problem of two dense tasks, i.e., semantic segment ation and depth estimation, and present a novel attention module called Cross-Ch annel Attention Module (CCAM), which facilitates effective feature sharing along each channel between the two tasks, leading to mutual performance gain with a n egligible increase in trainable parameters. In a symbiotic spirit, we also formu late novel data augmentations for the semantic segmentation task using predicted depth called AffineMix, and one using predicted semantics called ColorAug, for depth estimation task. Finally, we validate the performance gain of the proposed method on the Cityscapes and ScanNet dataset. which helps us achieve state-of-t he-art results for a semi-supervised joint model based on depth estimation and s emantic segmentation.
**********

Intra-Batch Supervision for Panoptic Segmentation on High-Resolution Images
Daan de Geus, Gijs Dubbelman; Proceedings of the IEEE/CVF Winter Conference on A pplications of Computer Vision (WACV), 2023, pp. 3165-3173
Unified panoptic segmentation methods are achieving state-of-the-art results on several datasets. To achieve these results on high-resolution datasets, these me thods apply crop-based training. In this work, we find that, although crop-based training is advantageous in general, it also has a harmful side-effect. Specifi cally, it limits the ability of unified networks to discriminate between large o bject instances, causing them to make predictions that are confused between mult iple instances. To solve this, we propose Intra-Batch Supervision (IBS), which i mproves a network's ability to discriminate between instances by introducing add itional supervision using multiple images from the same batch. We show that, wit h our IBS, we successfully address the confusion problem and consistently improv e the performance of unified networks. For the high-resolution Cityscapes and Ma pillary Vistas datasets, we achieve improvements of up to +2.5 on the Panoptic Q uality for thing classes, and even more considerable gains of up to +5.8 on both the pixel accuracy and pixel precision, which we identify as better metrics to capture the confusion problem.
**********

Temporally Consistent Online Depth Estimation in Dynamic Scenes
Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X. Creighton, Russell H. Taylor, Ganesh Venkatesh, Mathias Unberath; Proceedings of the IEEE/CVF Winter Conference on A pplications of Computer Vision (WACV), 2023, pp. 3018-3027
Temporally consistent depth estimation is crucial for online applications such a s augmented reality. While stereo depth estimation has received substantial atte ntion as a promising way to generate 3D information, there is relatively little work focused on maintaining temporal stability. Indeed, based on our analysis, c urrent techniques still suffer from poor temporal consistency. Stabilizing depth temporally in dynamic scenes is challenging due to concurrent object and camera motion. In an online setting, this process is further aggravated because only p ast frames are available. We present a framework named Consistent Online Dynamic Depth (CODD) to produce temporally consistent depth estimates in dynamic scenes in an online setting. CODD augments per-frame stereo networks with novel motion

and fusion networks. The motion network accounts for dynamics by predicting a per-pixel SE3 transformation and aligning the observations. The fusion network improves temporal depth consistency by aggregating the current and past estimates. We conduct extensive experiments and demonstrate quantitatively and qualitatively that CODD outperforms competing methods in terms of temporal consistency and performs on par in terms of per-frame accuracy.

**********

Hyperdimensional Feature Fusion for Out-of-Distribution Detection

Samuel Wilson, Tobias Fischer, Niko Sünderhauf, Feras Dayoub; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2644-2654

We introduce powerful ideas from Hyperdimensional Computing into the challenging field of Out-of-Distribution (OOD) detection. In contrast to many existing works that perform OOD detection based on only a single layer of a neural network, we use similarity-preserving semi-orthogonal projection matrices to project the feature maps from multiple layers into a common vector space. By repeatedly applying the bundling operation , we create expressive class-specific descriptor vectors for all in-distribution classes. At test time, a simple and efficient cosine similarity calculation between descriptor vectors consistently identifies OOD samples with competitive performance to the current state-of-the-art whilst being significantly faster. We show that our method is orthogonal to recent state-of-the-art OOD detectors and can be combined with them to further improve upon the performance.

**********

ALPINE: Improving Remote Heart Rate Estimation Using Contrastive Learning

Lokendra Birla, Sneha Shukla, Anup Kumar Gupta, Puneet Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5029-5038

Heart rate (HR) is a crucial physiological indicator of human health and can be used to detect cardiovascular disorders. The traditional HR estimation methods, such as electrocardiograms (ECG) and photoplethysmographs, require skin contact. Due to the increased risk of viral infection from skin contact, these approaches are avoided in the ongoing COVID-19 pandemic. Alternatively, one can use the non-contact HR estimation technique, remote photoplethysmography (rPPG), wherein HR is estimated from the facial videos of a person. Unfortunately, the existing rPPG methods perform poorly in the presence of facial deformations. Recently, there has been a proliferation of deep learning networks for rPPG. However, these networks require large-scale labelled data for better generalization. To alleviate these shortcomings, we propose a method ALPINE, that is, A noveL rPPG technique for Improving the remote heart rate estimatioN using contrastive lEarning. ALPINE utilizes the contrastive learning framework during training to address the issue of limited labelled data and introduces diversity in the data samples for better network generalization. Additionally, we introduce a novel hybrid loss comprising contrastive loss, signal-to-noise ratio (SNR) loss and data fidelity loss. Our novel contrastive loss maximizes the similarity between the rPPG information from different facial regions, thereby minimizing the effect of local noise. The SNR loss improves the quality of temporal signals, and the data fidelity loss ensures that the correct rPPG signal is extracted. Our extensive experiments on publicly available datasets demonstrate that the proposed method, ALPINE outperforms the previous well-known rPPG methods.

**********

Motif Mining: Finding and Summarizing Remixed Image Content

William Theisen, Daniel Gonzalez Cedre, Zachariah Carmichael, Daniel Moreira, Tim Weninger, Walter Scheirer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1319-1328

On the internet, images are no longer static; they have become dynamic content. Thanks to the availability of smartphones with cameras and easy-to-use editing software, images can be remixed (i.e., redacted, edited, and recombined with other content) on-the-fly and with a worldwide audience that can repeat the process. From digital art to memes, the evolution of images through time is now an impor

tant topic of study for digital humanists, social scientists, and media forensics specialists. However, because typical data sets in computer vision are composed of static content, the development of automated algorithms to analyze remixed content has been limited. In this paper, we introduce the idea of Motif Mining -- the process of finding and summarizing remixed image content in large collections of unlabeled and unsorted data. In this paper, this idea is formalized and a reference implementation is introduced. Experiments are conducted on three meme-style data sets, including a newly collected set associated with the information war in the Russo-Ukrainian conflict. The proposed motif mining approach is able to identify related remixed content that, when compared to similar approaches, more closely aligns with the preferences and expectations of human observers.
**********

SimGlim: Simplifying Glimpse Based Active Visual Reconstruction
Abhishek Jha, Soroush Seifi, Tinne Tuytelaars; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 269-278
An agent with a limited field of view needs to sample the most informative local observations of an environment in order to model the global context. Current works train this selection strategy by defining a complex architecture built upon features learned through convolutional encoders. In this paper, we first discuss why vision transformers are better suited than CNNs for such an agent. Next, we propose a simple transformer based active visual sampling model, called "SimGlim", which utilises transformer's inherent self-attention architecture to sequentially predict the best next location based on the current observable environment. We show the efficacy of our proposed method on the task of image reconstruction in the partial observable setting and compare our model against existing state-of-the-art active visual reconstruction methods. Finally, we provide ablations for the parameters of our design choice to understand their importance in the overall architecture.
**********

Treatment Learning Causal Transformer for Noisy Image Classification
Chao-Han Huck Yang, I-Te Hung, Yi-Chieh Liu, Pin-Yu Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6139-6150
Current top-notch deep learning (DL) based vision models are primarily based on exploring and exploiting the inherent correlations between training data samples and their associated labels. However, a known practical challenge is their degraded performance against "noisy" data, induced by different circumstances such as spurious correlations, irrelevant contexts, domain shift, and adversarial attacks. In this work, we incorporate this binary information of "existence of noise" as treatment into image classification tasks to improve prediction accuracy by jointly estimating their treatment effects. Motivated from causal variational inference, we propose a transformer-based architecture, Treatment Learning Causal Transformer (TLT), that uses a latent generative model to estimate robust feature representations from current observational input for noise image classification. Depending on the estimated noise level (modeled as a binary treatment factor), TLT assigns the corresponding inference network trained by the designed causal loss for prediction. We also create new noisy image datasets incorporating a wide range of noise factors (e.g., object masking, style transfer, and adversarial perturbation) for performance benchmarking. The superior performance of TLT in noisy image classification is further validated by several refutation evaluation metrics. As a by-product, TLT also improves visual salience methods for perceiving noisy images.
**********

OutfitTransformer: Learning Outfit Representations for Fashion Recommendation
Rohan Sarkar, Navaneeth Bodla, Mariya I. Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, Gerard Medioni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3601-3609
Learning an effective outfit-level representation is critical for predicting the compatibility of items in an outfit, and retrieving complementary items for a partial outfit. We present a framework, OutfitTransformer, that uses the proposed

task-specific tokens and leverages the self-attention mechanism to learn effective outfit-level representations encoding the compatibility relations between all items in the entire outfit for addressing both compatibility prediction and complementary item retrieval. For compatibility prediction, we design an outfit token to capture a global outfit representation and train the framework using a classification loss. For complementary item retrieval, we design a target item token that additionally takes the target item specification (in the form of a category or text description) into consideration. We train our framework using a proposed set-wise outfit ranking loss to generate a target item embedding given an outfit, and a target item specification as inputs. The generated target item embedding is then used to retrieve compatible items that match the rest of the outfit. Additionally, we adopt a pre-training approach and a curriculum learning strategy to improve retrieval performance. Experiments show that our approach outperforms state-of-the-art methods on compatibility prediction, fill-in-the-blank, and complementary item retrieval tasks.

**********

No Shifted Augmentations (NSA): Compact Distributions for Robust Self-Supervised Anomaly Detection

Mohamed Yousef, Marcel Ackermann, Unmesh Kurup, Tom Bishop; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5511-5520

Unsupervised Anomaly detection (AD) requires building a notion of normalcy, distinguishing in-distribution (ID) and out-of-distribution (OOD) data, using only available ID samples. Recently, large gains were made on this task for the domain of natural images using self-supervised contrastive feature learning as a first step followed by kNN or traditional one-class classifiers for feature scoring. Learned representations that are non-uniformly distributed on the unit hypersphere have been shown to be beneficial for this task. We go a step further and investigate how the ph geometrical compactness of the ID feature distribution makes isolating and detecting outliers easier, especially in the realistic situation when ID training data is polluted with OOD data. We propose novel architectural modifications to the self-supervised feature learning step, that enable such compact distributions for ID data to be learned. We show that the proposed modifications can be effectively applied to most existing self-supervised objectives, with large gains in performance. Furthermore, this improved OOD performance is obtained without resorting to tricks such as using strongly augmented ID images (e.g. by 90 degree rotations) as proxies for the unseen OOD data, as these impose overly prescriptive assumptions about ID data and its invariances. We perform extensive studies on benchmark datasets for one-class OOD detection and show state-of-the-art performance in the presence of pollution in the ID data, and comparable performance otherwise. We also propose and extensively evaluate a novel feature scoring technique based on the angular Mahalanobis distance, and propose a simple and novel technique for feature ensembling during evaluation that enables a big boost in performance at nearly zero run-time cost compared to the standard use of model ensembling or test time augmentations. Source code is available here: https://github.com/IntuitionMachines/NSA

**********

I See-Through You: A Framework for Removing Foreground Occlusion in Both Sparse and Dense Light Field Images

Jiwan Hur, Jae Young Lee, Jaehyun Choi, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 229-238

Light field (LF) camera captures rich information from a scene. Using the information, the LF de-occlusion (LF-DeOcc) task aims to reconstruct the occlusion-free center view image. Existing LF-DeOcc studies mainly focus on the sparsely sampled (sparse) LF images where most of the occluded regions are visible in other views due to the large disparity. In this paper, we expand LF-DeOcc in more challenging datasets, densely sampled (dense) LF images, which are taken by a micro-lens-based portable LF camera. Due to the small disparity ranges of dense LF images, most of the background regions are invisible in any view. To apply LF-DeOcc in both LF datasets, we propose a framework, ISTY, which is defined and divided

into three roles: (1) extract LF features, (2) define the occlusion, and (3) inp aint occluded regions. By dividing the framework into three specialized componen ts according to the roles, the development and analysis can be easier. Furthermo re, an explainable intermediate representation, an occlusion mask, can be obtain ed in the proposed framework. The occlusion mask is useful for comprehensive ana lysis of the model and other applications by manipulating the mask. In experimen ts, qualitative and quantitative results show that the proposed framework outper forms state-of-the-art LF-DeOcc methods in both sparse and dense LF datasets.
**********

Generative Colorization of Structured Mobile Web Pages
Kotaro Kikuchi, Naoto Inoue, Mayu Otani, Edgar Simo-Serra, Kota Yamaguchi; Proce edings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC V), 2023, pp. 3650-3659
Color is a critical design factor for web pages, affecting important factors suc h as viewer emotions and the overall trust and satisfaction of a website. Effect ive coloring requires design knowledge and expertise, but if this process could be automated through data-driven modeling, efficient exploration and alternative workflows would be possible. However, this direction remains underexplored due to the lack of a formalization of the web page colorization problem, datasets, a nd evaluation protocols. In this work, we propose a new dataset consisting of e-commerce mobile web pages in a tractable format, which are created by simplifyin g the pages and extracting canonical color styles with a common web browser. The web page colorization problem is then formalized as a task of estimating plausi ble color styles for a given web page content with a given hierarchical structur e of the elements. We present several Transformer-based methods that are adapted to this task by prepending structural message passing to capture hierarchical r elationships between elements. Experimental results, including a quantitative ev aluation designed for this task, demonstrate the advantages of our methods over statistical and image colorization methods. The code is available at https://git hub.com/CyberAgentAILab/webcolor.
**********

Style-Guided Inference of Transformer for High-Resolution Image Synthesis
Jonghwa Yim, Minjae Kim; Proceedings of the IEEE/CVF Winter Conference on Applic ations of Computer Vision (WACV), 2023, pp. 1745-1755
Transformer is eminently suitable for auto-regressive image synthesis which pred icts discrete value from the past values recursively to make up full image. Espe cially, combined with vector quantised latent representation, the state-of-the-a rt auto-regressive transformer displays realistic high-resolution images. Howeve r, sampling the latent code from discrete probability distribution makes the out put unpredictable. Therefore, it requires to generate lots of diverse samples to acquire desired outputs. To alleviate the process of generating lots of samples repetitively, in this article, we propose to take a desired output, a style ima ge, as an additional condition without re-training the transformer. To this end, our method transfers the style to a probability constraint to re-balance the pr ior, thereby specifying the target distribution instead of the original prior. T hus, generated samples from the re-balanced prior have similar styles to referen ce style. In practice, we can choose either an image or a category of images as an additional condition. In our qualitative assessment, we show that styles of m ajority of outputs are similar to the input style.
**********

Resolving Class Imbalance for LiDAR-Based Object Detector by Dynamic Weight Aver age and Contextual Ground Truth Sampling
Daeun Lee, Jinkyu Kim; Proceedings of the IEEE/CVF Winter Conference on Applicat ions of Computer Vision (WACV), 2023, pp. 682-691
An autonomous driving system requires a 3D object detector, which must perceive all present road agents reliably to navigate an environment safely. However, rea l-world driving datasets often suffer from the problem of data imbalance, which causes difficulties in training a model that works well across all classes, resu lting in an undesired imbalanced sub-optimal performance. In this work, we propo se a method to address this data imbalance problem. Our method consists of two m

ain components: (i) a LiDAR-based 3D object detector with per-class multiple det ection heads where losses from each head are modified by dynamic weight average to be balanced. (ii) Contextual ground truth (GT) sampling, where we improve con ventional GT sampling techniques by leveraging semantic information to augment p oint cloud with sampled ground truth GT objects. Our experiment with KITTI and n uScenes datasets confirms our proposed method's effectiveness in dealing with th e data imbalance problem, producing better detection accuracy compared to existi ng approaches. Our implementation will be publicly available upon publication.
**********

One-Shot Doc Snippet Detection: Powering Search in Document Beyond Text
Abhinav Java, Shripad Deshmukh, Milan Aggarwal, Surgan Jandial, Mausoom Sarkar, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applicati ons of Computer Vision (WACV), 2023, pp. 5437-5446
Active consumption of digital documents has yielded scope for research in variou s applications, including search. Traditionally, searching within a document has been cast as a text matching problem ignoring the rich layout and visual cues c ommonly present in structured documents, forms, etc. To that end, we ask a mostl y unexplored question: "Can we search for other similar snippets present in a ta rget document page given a single query instance of a document snippet?". We pro pose MONOMER to solve this as a one-shot snippet detection task. MONOMER fuses c ontext from visual, textual, and spatial modalities of snippets and documents to find query snippet in target documents. We conduct extensive ablations and expe riments showing MONOMER outperforms several baselines from one-shot object detec tion (BHRL), template matching, and document understanding (LayoutLMv3). Due to the scarcity of relevant data for the task at hand, we train MONOMER on programm atically generated data having many visually similar query snippets and target d ocument pairs from two datasets - Flamingo Forms and PubLayNet. We also do a hum an study to validate the generated data.
**********

Meta-OLE: Meta-Learned Orthogonal Low-Rank Embedding
Ze Wang, Yue Lu, Qiang Qiu; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 5305-5314
We introduce Meta-OLE, a new geometry-regularized method for fast adaptation to novel tasks in few-shot image classification. The proposed method learns to adap t for each few-shot classification task a feature space with simultaneous inter- class orthogonality and intra-class low-rankness. Specifically, a deep feature e xtractor is trained by explicitly imposing orthogonal low-rank subspace structur es among features corresponding to different classes within a given task. To ada pt to novel tasks with unseen categories, we further meta-learn a light-weight t ransformation to enhance the inter-class margins. As an additional benefit, this light-weight transformation lets us exploit the query data for label propagatio n from labeled to unlabeled data without any auxiliary network components. The e xplicitly geometry-regularized feature subspaces allow the classifiers on novel tasks to be inferred in a closed form, with an adaptive subspace truncation that selectively discards non-discriminative dimensions. We perform experiments on s tandard few-shot image classification tasks, and observe performance superior to state-of-the-art meta-learning methods.
**********

Few-Shot Object Detection via Improved Classification Features
Xinyu Jiang, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, Duoqian Miao; Proc eedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WA CV), 2023, pp. 5386-5395
Few-shot object detection (FSOD) aims to transfer knowledge from base classes to novel classes, which receives widespread attention recently. The performance of current techniques is, however, limited by the poor classification ability and the improper features in the detection head. To circumvent this issue, we propos e a Multi-level Feature Enhancement (MFE) model to improve the feature for class ification from three different perspectives, including the spatial level, the ta sk level and the regularization level. First, we revise the classifier's input f eature at the spatial level by using information from the regression head. Secon

dly, we separate the RoI-Align feature into two different feature distributions in order to improve features at the task level. Finally, taking into account the overfitting problem in FSOD, we design a simple but efficient regularization enhancement module to sample features into various distributions and enhance the regularization ability of classification. Extensive experiments show that our method achieves competitive results on PASCAL VOC datasets, and exceeds current state-of-the-art methods in all shot settings on challenging MS-COCO datasets.
**********

WSNet: Towards an Effective Method for Wound Image Segmentation
Subba Reddy Oota, Vijay Rowtula, Shahid Mohammed, Minghsun Liu, Manish Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3234-3243
Medical image segmentation is critical for effective computer-aided diagnosis and localization of ailments. Automated segmentation of wound regions from patient images can aid clinicians in measuring and managing chronic wounds and monitoring the wound healing trajectory. While there exists a plethora of work on general medical image segmentation, there is hardly any work on wound image analysis and segmentation. Existing methods are limited to segmenting a smaller subset of ulcers, such as foot ulcers, with no special processing for wound images. In this paper, we build segmentation models for eight different types of wound images. Wound image analysis is a challenging problem due to the lack of availability of extensive data (labeled or unlabeled) and annotation challenges due to the shortage of well-trained wound care clinicians. To handle these challenges, we contribute WoundSeg, a large and diverse dataset of segmented wound images. Generic wound image segmentation is complex due to the heterogeneous appearance of wound area across images of similar wound types. We propose a novel image segmentation framework, WSNet, which leverages (a) wound-domain adaptive pretraining on a large unlabeled wound image collection and (b) a global-local architecture that utilizes full image and its patches to learn fine-grained details of heterogeneous wounds. On WoundSeg, we achieve a decent Dice score of 0.847. On existing AZH Woundcare and Medetec datasets, we establish a new state-of-the-art. Further, we show the impact of using segmentation for improving the accuracy of downstream tasks like wound area and volume prediction.
**********

Image-Free Domain Generalization via CLIP for 3D Hand Pose Estimation
Seongyeong Lee, Hansoo Park, Dong Uk Kim, Jihyeon Kim, Muhammadjon Boboev, Seungryul Baek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2934-2944
RGB-based 3D hand pose estimation has been successful for decades thanks to large-scale databases and deep learning. However, the hand pose estimation network does not operate well for hand pose images whose characteristics are far different from the training data. This is caused by various factors such as illuminations, camera angles, diverse backgrounds in the input images, etc. Many existing methods tried to solve it by supplying additional large-scale unconstrained/target domain images to augment data space; however collecting such large-scale images takes a lot of labors. In this paper, we present a simple image-free domain generalization approach for the hand pose estimation framework that uses only source domain data. We try to manipulate the image features of the hand pose estimation network by adding the features from text descriptions using the CLIP (Contrastive Language-Image Pre-training) model. The manipulated image features are then exploited to train the hand pose estimation network via the contrastive learning framework. In experiments with STB and RHD datasets, our algorithm shows improved performance over the state-of-the-art domain generalization approaches.
**********

TVCalib: Camera Calibration for Sports Field Registration in Soccer
Jonas Theiner, Ralph Ewerth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1166-1175
Sports field registration in broadcast videos is typically interpreted as the task of homography estimation, which provides a mapping between a planar field and the corresponding visible area of the image. In contrast to previous approaches

, we consider the task as a camera calibration problem. First, we introduce a differentiable objective function that is able to learn the camera pose and focal length from segment correspondences (e.g., lines, point clouds), based on pixel-level annotations for segments of a known calibration object. The calibration module iteratively minimizes the segment reprojection error induced by the estimated camera parameters. Second, we propose a novel approach for 3D sports field registration from broadcast soccer images. Compared to the typical solution, which subsequently refines an initial estimation, our solution does it in one step. The proposed method is evaluated for sports field registration on two datasets and achieves superior results compared to two state-of-the-art approaches.
**********

Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution
Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, Tae Hyun Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4956-4965
Recent transformer-based super-resolution (SR) methods have achieved promising results against conventional CNN-based methods. However, these approaches suffer from essential shortsightedness created by only utilizing the standard self-attention-based reasoning. In this paper, we introduce an effective hybrid SR network to aggregate enriched features, including local features from CNNs and long-range multi-scale dependencies captured by transformers. Specifically, our network comprises transformer and convolutional branches, which synergetically complement each representation during the restoration procedure. Furthermore, we propose a cross-scale token attention module, allowing the transformer branch to exploit the informative relationships among tokens across different scales efficiently. Our proposed method achieves state-of-the-art SR results on numerous benchmark datasets.
**********

Perceptual Image Enhancement for Smartphone Real-Time Applications
Marcos V. Conde, Florin Vasluianu, Javier Vazquez-Corral, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1848-1858
Recent advances in camera designs and imaging pipelines allow us to capture high-quality images using smartphones. However, due to the small size and lens limitations of the smartphone cameras, we commonly find artifacts or degradation in the processed images. The most common unpleasant effects are noise artifacts, diffraction artifacts, blur, and HDR overexposure. Deep learning methods for image restoration can successfully remove these artifacts. However, most approaches are not suitable for real-time applications on mobile devices due to their heavy computation and memory requirements. In this paper, we propose LPIENet, a lightweight network for perceptual image enhancement, with the focus on deploying it on smartphones. Our experiments show that, with much fewer parameters and operations, our model can deal with the mentioned artifacts and achieve competitive performance compared with state-of-the-art methods on standard benchmarks. Moreover, to prove the efficiency and reliability of our approach, we deployed the model directly on commercial smartphones and evaluated its performance. Our model can process 2K resolution images under 1s without specific optimization for the mobile devices.
**********

CountNet3D: A 3D Computer Vision Approach To Infer Counts of Occluded Objects
Porter Jenkins, Kyle Armstrong, Stephen Nelson, Siddhesh Gotad, J. Stockton Jenkins, Wade Wilkey, Tanner Watts; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3008-3017
3D scene understanding is an important problem that has experienced great progress in recent years, in large part due to the development of state-of-the-art methods for 3D object detection. However, the performance of 3D object detectors can suffer in scenarios where extreme occlusion of objects is present, or the number of object classes is large. In this paper, we study the problem of inferring 3D counts from densely packed scenes with heterogeneous objects. This problem has applications to important tasks such as inventory management or automatic crop

yield estimation. We propose a novel regression-based method, CountNet3D, that uses mature 2D object detectors for finegrained classification and localization, and a PointNet backbone for geometric embedding. The network processes fused data from images and point clouds for end-to-end learning of counts. We perform experiments on a novel synthetic dataset for inventory management in retail, which we construct and make publicly available to the community. Our results show that regression-based 3D counting methods systematically outperform detection-based methods, and reveal that directly learning from raw point clouds greatly assists count estimation under extreme occlusion. Finally, we study the effectiveness of CountNet3D on a large dataset of real-world scenes where extreme occlusion is present and achieve an error rate of 11.01%.

**********

The Box Size Confidence Bias Harms Your Object Detector

Johannes Gilg, Torben Teepe, Fabian Herzog, Gerhard Rigoll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1471-1480

Countless applications depend on accurate predictions with reliable confidence estimates from modern object detectors. However, it is well known that neural networks, including object detectors, produce miscalibrated confidence estimates. Recent work even suggests that detectors' confidence predictions are biased with respect to object size and position. In object detection, the issues of conditional biases, confidence calibration, and task performance are usually explored in isolation, but, as we aim to show, they are closely related. We formally prove that the conditional confidence bias harms the performance of object detectors and empirically validate these findings. Specifically, to quantify the performance impact of the confidence bias on object detectors, we modify the histogram binning calibration to avoid performance impairment and instead improve it through calibration conditioned on the bounding box size. We further find that the confidence bias is also present in detections generated on the training data of the detector, which can be leveraged to perform the de-biasing. Moreover, we show that Test Time Augmentation (TTA) confounds this bias, which results in even more significant performance impairments on the detectors. Finally, we use our proposed algorithm to analyze a diverse set of object detection architectures and show that the conditional confidence bias harms their performance by up to 0.6 mAP and 0.8 mAP50. Code available at https://github.com/Blueblue4/Object-Detection-Confidence-Bias.

**********

D2F2WOD: Learning Object Proposals for Weakly-Supervised Object Detection via Progressive Domain Adaptation

Yuting Wang, Ricardo Guerrero, Vladimir Pavlovic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 22-31

Weakly-supervised object detection (WSOD) models attempt to leverage image-level annotations in lieu of accurate but costly-to-obtain object localization labels. This oftentimes leads to substandard object detection and localization at inference time. To tackle this issue, we propose D2DF2WOD, a Dual-Domain Fully-to-Weakly Supervised Object Detection framework that leverages synthetic data, annotated with precise object localization, to supplement a natural image target domain, where only image-level labels are available. In its warm-up domain adaptation stage, the model learns a fully-supervised object detector (FSOD) to improve the precision of the object proposals in the target domain, and at the same time learns target-domain-specific and detection-aware proposal features. In its main WSOD stage, a WSOD model is specifically tuned to the target domain. The feature extractor and the object proposal generator of the WSOD model are built upon the fine-tuned FSOD model. We test D2DF2WOD on five dual-domain image benchmarks. The results show that our method results in consistently improved object detection and localization compared with state-of-the-art methods.

**********

More Control for Free! Image Synthesis With Semantic Diffusion Guidance

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, Trevor Darrell; Proceedings of the IEEE/CVF Winter

Conference on Applications of Computer Vision (WACV), 2023, pp. 289-299
Controllable image synthesis models allow creation of diverse images based on text instructions or guidance from a reference image. Recently, denoising diffusion probabilistic models have been shown to generate more realistic imagery than prior methods, and have been successfully demonstrated in unconditional and class-conditional settings. We investigate fine-grained, continuous control of this model class, and introduce a novel unified framework for semantic diffusion guidance, which allows either language or image guidance, or both. Guidance is injected into a pretrained unconditional diffusion model using the gradient of image-text or image matching scores, without re-training the diffusion model. We explore CLIP-based language guidance as well as both content and style-based image guidance in a unified framework. Our text-guided synthesis approach can be applied to datasets without associated text annotations. We conduct experiments on FFHQ and LSUN datasets, and show results on fine-grained text-guided image synthesis, synthesis of images related to a style or content reference image, and examples with both textual and image guidance.

**********

Lossy Image Compression With Quantized Hierarchical VAEs
Zhihao Duan, Ming Lu, Zhan Ma, Fengqing Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 198-207
Recent work has shown a strong theoretical connection between variational autoencoders (VAEs) and the rate distortion theory. Motivated by this, we consider the problem of lossy image compression from the perspective of generative modeling. Starting from ResNet VAEs, which are originally designed for data (image) distribution modeling, we redesign their latent variable model using a quantization-aware posterior and prior, enabling easy quantization and entropy coding for image compression. Along with improved neural network blocks, we present a powerful and efficient class of lossy image coders, outperforming previous methods on natural image (lossy) compression. Our model compresses images in a coarse-to-fine fashion and supports parallel encoding and decoding, leading to fast execution on GPUs.

**********

HyperShot: Few-Shot Learning by Kernel HyperNetworks
Marcin Sendera, Marcin Przewi■■likowski, Konrad Karanowski, Maciej Zi■ba, Jacek Tabor, Przemys■aw Spurek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2469-2478
Few-shot models aim at making predictions using a minimal number of labeled examples from a given task. The main challenge in this area is the one-shot setting where only one element represents each class. We propose HyperShot - the fusion of kernels and hypernetwork paradigm. Compared to reference approaches that apply a gradient-based adjustment of the parameters, our model aims to switch the classification module parameters depending on the task's embedding. In practice, we utilize a hypernetwork, which takes the aggregated information from support data and returns the classifier's parameters handcrafted for the considered problem. Moreover, we introduce the kernel-based representation of the support examples delivered to hypernetwork to create the parameters of the classification module. Consequently, we rely on relations between embeddings of the support examples instead of direct feature values provided by the backbone models. Thanks to this approach, our model can adapt to highly different tasks.

**********

MMPTRACK: Large-Scale Densely Annotated Multi-Camera Multiple People Tracking Benchmark
Xiaotian Han, Quanzeng You, Chunyu Wang, Zhizheng Zhang, Peng Chu, Houdong Hu, Jiang Wang, Zicheng Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4860-4869
Multi-camera tracking systems are gaining popularity in applications that demand high-quality tracking results, such as frictionless checkout. In cluttered and crowded environments, monocular multi-object tracking (MOT) systems often fail due to occlusions. Multiple highly overlapped cameras are capable of recovering partial 3D information. When used properly, 3D data can significantly alleviate t

he occlusion issue. However, training a multi-camera tracker demands a large-sca
le multi-camera tracking dataset with diverse camera settings and backgrounds. T
hese requirements make the collection of multi-camera tracking dataset challengi
ng and expensive. The cost of creating such a dataset has limited the availabili
ty and scale of datasets in this domain. Instead, we appeal to an auto-annotatio
n system to reduce the cost. The system uses overlapped and calibrated depth and
 RGB cameras to build a 3D tracker and automatically generates the 3D tracking r
esults. We then manually check and correct the 3D tracking results to ensure the
 label quality, which is much cheaper than solely manual annotation. Next, the 3
D tracking results are projected to each calibrated RGB camera view to create 2D
 tracking results. In this way, we collect and annotate a large-scale densely la
beled multi-camera tracking dataset from five different environments. We have co
nducted extensive experiments using two real-time multi-camera trackers and a pe
rson re-identification (ReID) model under different settings. This dataset provi
des a reliable benchmark for multi-camera, multi-object tracking systems in clut
tered and crowded environments. We expect this benchmark to encourage more resea
rch attempts in this domain. Also, our results demonstrate that adapting the tra
ckers and ReID models on this dataset significantly improves their performance.
Our dataset will be publicly released upon the acceptance of this work.
**********

SHARDS: Efficient Shadow Removal Using Dual Stage Network for High-Resolution Im
ages

Mrinmoy Sen, Sai Pradyumna Chermala, Nazrinbanu Nurmohammad Nagori, Venkat Peddi
gari, Praful Mathur, B. H. Pawan Prasad, Moonhwan Jeong; Proceedings of the IEEE
/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1809
-1817

Shadow Removal is an important and widely researched topic in computer vision. R
ecent advances in deep learning have resulted in addressing this problem by usin
g convolutional neural networks (CNNs) similar to other vision tasks. But these
existing works are limited to low-resolution images. Furthermore, the existing m
ethods rely on heavy network architectures which cannot be deployed on resource-
constrained platforms like smartphones. In this paper, we propose SHARDS, a shad
ow removal method for high-resolution images. The proposed method solves shadow
removal for high-resolution images in two stages using two lightweight networks:
 a Low-resolution Shadow Removal Network (LSRNet) followed by a Detail Refinemen
t Network (DRNet). LSRNet operates at low-resolution and computes a low-resoluti
on, shadow-free output. It achieves state-of-the-art results on standard dataset
s with 65x lesser network parameters than existing methods. This is followed by
DRNet, which is tasked to refine the low-resolution output to a high-resolution
output using the high-resolution input shadow image as guidance. We construct hi
gh-resolution shadow removal datasets and through our experiments, prove the eff
ectiveness of our proposed method on them. It is then demonstrated that this met
hod can be deployed on modern day smartphones and is the first of its kind solut
ion that can efficiently (2.4secs) perform shadow removal for high-resolution im
ages (12MP) in these devices. Like many existing approaches, our shadow removal
network relies on a shadow region mask as input to the network. To complement th
e lightweight shadow removal network, we also propose a lightweight shadow detec
tor in this paper.
**********

Robustness of Trajectory Prediction Models Under Map-Based Attacks

Zhihao Zheng, Xiaowen Ying, Zhen Yao, Mooi Choo Chuah; Proceedings of the IEEE/C
VF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4541-4
550

Trajectory Prediction (TP) is a critical component in the control system of an A
utonomous Vehicle (AV). It predicts future motion of traffic agents based on obs
ervations of their past trajectories. Existing works have studied the vulnerabil
ity of TP models when the perception systems are under attacks and proposed corr
esponding mitigation schemes. Recent TP designs have incorporated context map in
formation for performance enhancements. Such designs are subjected to a new type
 of attacks where an attacker can interfere with these TP models by attacking th

e context maps. In this paper, we study the robustness of TP models under our ne wly proposed map-based adversarial attacks. We show that such attacks can compro mise state-of-the-art TP models that use either image-based or node-based map re presentation while keeping the adversarial examples imperceptible. We also demon strate that our attacks can still be launched under the black-box settings witho ut any knowledge of the TP models running underneath. Our experiments on the NuS cene dataset show that the proposed map-based attacks can increase the trajector y prediction errors by 29-110%. Finally, we demonstrate that two defense mechani sms are effective in defending against such map-based attacks.
**********

Computer Vision for International Border Legibility
Trevor Ortega, Thomas Nelson, Skyler Crane, Josh Myers-Dean, Scott Wehrwein; Pro ceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (W ACV), 2023, pp. 3838-3847
Key aspects of international policy, such as those pertaining to migration and t rade, manifest in the physical world at international political borders; for thi s reason, borders are of interest to political science studying the impacts and implications of these policies. While some prior efforts have worked to characte rize features of borders using trained human coders and crowdsourcing, these are limited in scale by the need for manual annotations. In this paper, we present a new task, dataset, and baseline approaches for estimating the legibility of in ternational political borders automatically and on a global scale. Our contribut ions are to (1) define the border legibility estimation task; (2) collect a data set of overhead (aerial) imagery for the entire world's international borders, ( 3) propose several classical and deep-learning-based approaches to establish a b aseline for the task, and (4) evaluate our algorithms against a validation datas et of crowdsourced legibility comparisons. Our results on this challenging task confirm that while low-level features can often explain border legibility, mid- and high-level features are also important. Finally, we show preliminary results of a global analysis of legibility, confirming some of the political and geogra phic influences of legibility.
**********

Fractual Projection Forest: Fast and Explainable Point Cloud Classifier
Hanxiao Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Co mputer Vision (WACV), 2023, pp. 4240-4249
Point clouds are playing an increasingly important roll in autonomous driving an d robotics. Although current point cloud classification models have achieved sat isfactory accuracies, most of them trade slight performance gains by stacking co mplex modules on the grouping-local-global framework, which leads to prolonged p rocessing time and deteriorating interpretability. In this work, we propose a ne w pipeline named Fractual Projection Forest (FPF) that exploits fractal features to enable traditional machine learning models to achieve competitive performanc e with DNNs on classification tasks. Though compromises by few percentages in ac curacy compared to DNNs, FPF is faster, more interpretable, and easily extendabl e. We hope that FPF may provide the community with a novel view of point cloud c lassification. Our code is available on https://github.com/Explain3D/FracProjFor est.
**********

Improving Deep Facial Phenotyping for Ultra-Rare Disorder Verification Using Mod el Ensembles
Alexander Hustinx, Fabio Hellmann, Ömer Sümer, Behnam Javanmardi, Elisabeth Andr é, Peter Krawitz, Tzung-Chien Hsieh; Proceedings of the IEEE/CVF Winter Conferen ce on Applications of Computer Vision (WACV), 2023, pp. 5018-5028
Rare genetic disorders affect more than 6% of the global population. Reaching a diagnosis is challenging because rare disorders are very diverse. Many disorders have recognizable facial features that are hints for clinicians to diagnose pat ients. Previous work, such as GestaltMatcher, utilized representation vectors pr oduced by a DCNN similar to AlexNet to match patients in high-dimensional featur e space to support "unseen" ultra-rare disorders. However, the architecture and dataset used for transfer learning in GestaltMatcher have become outdated. Moreo

ver, a way to train the model for generating better representation vectors for u nseen ultra-rare disorders has not yet been studied. Because of the overall scar city of patients with ultra-rare disorders, it is infeasible to directly train a model on them. Therefore, we first analyzed the influence of replacing GestaltM atcher DCNN with a state-of-the-art face recognition approach, iResNet with ArcF ace. Additionally, we experimented with different face recognition datasets for transfer learning. Furthermore, we proposed test-time augmentation, and model en sembles that mix general face verification models and models specific for verify ing disorders to improve the disorder verification accuracy of unseen ultra-rare disorders. Our proposed ensemble model achieves state-of-the-art performance on both seen and unseen disorders. Code is available at https://www.github.com/igs b/GestaltMatcher-Arc

**********

Complementary Cues From Audio Help Combat Noise in Weakly-Supervised Object Dete ction

Cagri Gungor, Adriana Kovashka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2185-2194

We tackle the problem of learning object detectors in a noisy environment, which is one of the significant challenges for weakly-supervised learning. We use mul timodal learning to help localize objects of interest, but unlike other methods, we treat audio as an auxiliary modality that assists to tackle noise in detecti on from visual regions. First, we use the audio-visual model to generate new "gr ound-truth" labels for the training set to remove noise between the visual featu res and noisy supervision. Second, we propose an "indirect path" between audio a nd class predictions, which combines the link between visual and audio regions, and the link between visual features and predictions. Third, we propose a sound-based "attention path" which uses the benefit of complementary audio cues to ide ntify important visual regions. We use contrastive learning to perform region-ba sed audio-visual instance discrimination, which serves as an intermediate task a nd benefits from the complementary cues from audio to boost object classificatio n and detection performance. We show that our methods, which update noisy ground truth and provide indirect and attention paths, greatly boosting performance on the AudioSet and VGGSound datasets compared to single-modality predictions, eve n ones that use contrastive learning. Our method outperforms previous weakly-sup ervised detectors for the task of object detection by reaching the state-of-art on AudioSet, and our sound localization module performs better than several stat e-of-art methods on AudioSet and MUSIC.

**********

One-Shot Synthesis of Images and Segmentation Masks

Vadim Sushko, Dan Zhang, Jürgen Gall, Anna Khoreva; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6285-6294

Joint synthesis of images and segmentation masks with generative adversarial net works (GANs) is promising to reduce the effort needed for collecting image data with pixel-wise annotations. However, to learn high-fidelity image-mask synthesi s, existing GAN approaches first need a pre-training phase requiring large amoun ts of image data, which limits their utilization in restricted image domains. In this work, we take a step to reduce this limitation, introducing the task of on e-shot image-mask synthesis. We aim to generate diverse images and their segment ation masks given only a single labelled example, and assuming, contrary to prev ious models, no access to any pre-training data. To this end, inspired by the re cent architectural developments of single-image GANs, we introduce our OSMIS mod el which enables the synthesis of segmentation masks that are precisely aligned to the generated images in the one-shot regime. Besides achieving the high fidel ity of generated masks, OSMIS outperforms state-of-the-art single-image GAN mode ls in image synthesis quality and diversity. In addition, despite not using any additional data, OSMIS demonstrates an impressive ability to serve as a source o f useful data augmentation for one-shot segmentation applications, providing per formance gains that are complementary to standard data augmentation techniques. Code is available at https://github.com/boschresearch/one-shot-synthesis.

**********

Representation Disentanglement in Generative Models With Contrastive Learning
Shentong Mo, Zhun Sun, Chao Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1531-1540
Contrastive learning has shown its effectiveness in image classification and generation. Recent works apply the contrastive learning on the discriminator of the Generative Adversarial Networks, and there exists little work on exploring if contrastive learning can be applied on encoders to learn disentangled representations. In this work, we propose a simple yet effective method via incorporating contrastive learning into latent optimization, where we name it. Specifically, we first use a generator to learn discriminative and disentangled embeddings via latent optimization. Then an encoder and two momentum encoders are applied to dynamically learn disentangled information across large amount of samples with content-level and residual-level contrastive loss. In the meanwhile, we tune the encoder with the learned embeddings in an amortized manner. We evaluate our approach on ten benchmarks in terms of representation disentanglement and linear classification. Extensive experiments demonstrate the effectiveness of our ContraLORD on learning both discriminative and generative representations.
**********
Improving the Pair Selection and the Model Fusion Steps of Satellite Multi-View Stereo Pipelines
Alvaro Gómez, Gregory Randall, Gabriele Facciolo, Rafael Grompone von Gioi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6344-6353
Multi-view stereo reconstruction of scenes from satellite images is traditionally performed with a pair-wise stereo-vision approach: (1) multiple views are grouped into pairs, (2) each pair is processed by two-view stereo methods producing an elevation model or point cloud, lastly (3) the pair-wise reconstructions are integrated and filtered to obtain a final result. These steps are organized in a pipeline and the end-to-end performance of reconstructions depends on the behavior of these steps. This work introduces two changes that increase the performance of the reconstructions: a new pair selection approach and a new integration method are presented. The new pair selection replaces commonly used heuristics with a principled criterion that predicts the completeness of a pair based on offline simulations. The presented integration method is based on an iterated bilateral filter. Experiments show that these changes yield a systematic improvement on the performance of the pipeline.
**********
Learning Style Subspaces for Controllable Unpaired Domain Translation
Gaurav Bhatt, Vineeth N. Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4220-4229
The unpaired domain-to-domain translation aims to learn inter-domain relationships between diverse modalities without relying on paired data, which can help complex structure prediction tasks such as age transformation, where it is challenging to attain paired samples. A common approach used by most current methods is to factorize the data into a domain-invariant content space and a domain-specific style space. In this work, we argue that the style space can be further decomposed into smaller subspaces. Learning these style subspaces has two-fold advantages: (i) it allows more robustness and reliability in the generation of images in unpaired domain translation; and (ii) it allows better control and thereby interpolating the latent space, which can be helpful in complex translation tasks involving multiple domains. To achieve this decomposition, we propose a novel scalable approach to partition the latent space into style subspaces. We also propose a new evaluation metric that quantifies the controllable generation capability of domain translation methods. We compare our proposed method with several strong baselines on standard domain translation tasks such as gender translation (male-to-female and female-to-male), age transformation, reference-guided image synthesis, multi-domain image translation, and multi-attribute domain translation on celebA-HQ and AFHQ datasets. The proposed technique achieves state-of-the-art performance on various domain translation tasks while outperforming all the baselines on controllable generation tasks.

**********
NAPReg: Nouns As Proxies Regularization for Semantically Aware Cross-Modal Embeddings
Bhavin Jawade, Deen Dayal Mohan, Naji Mohamed Ali, Srirangaraj Setlur, Venu Govindaraju; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1135-1144
Cross-Modal retrieval is a fundamental vision-language task with a broad range of practical applications. Text-to-image matching is the most common form of cross-modal retrieval where given a large database of images and a textual query, the task is to retrieve the most relevant set of images. Existing methods utilize dual encoders with an attention mechanism and a ranking loss for learning embeddings that can be used for retrieval based on cosine similarity. Despite the fact that existing methods attempt to perform semantic alignment across visual regions and textual words using tailored attention mechanisms, there is no explicit supervision from the training objective to enforce such alignment. To address this, we propose NAPReg, a novel regularization formulation that projects high-level semantic entities i.e Nouns into the embedding space as shared learnable proxies. We show that using such a formulation allows the attention mechanism to learn better word-region alignment while also utilizing region information from other samples to build a more generalized latent representation for semantic concepts. Experiments on three benchmark datasets i.e. MS-COCO, Flickr30k and Flickr8k demonstrates that our method achieves state-of-the-art results in cross-modal metric learning for text-image and image-text retrieval tasks.
**********
Complementary Bi-Directional Feature Compression for Indoor 360deg Semantic Segmentation With Self-Distillation
Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, Yao Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4501-4510
Semantic segmentation on 360deg images is a vital component of scene understanding due to the rich surrounding information. Recently, horizontal representation-based approaches outperform projection-based solutions, because the distortions can be effectively removed by compressing the spherical data in the vertical direction. However, these methods ignore the distortion distribution prior and are limited to unbalanced receptive fields, e.g., the receptive fields are sufficient in the vertical direction and insufficient in the horizontal direction. Differently, a vertical representation compressed in another direction can offer implicit distortion prior and enlarge horizontal receptive fields. In this paper, we combine the two different representations and propose a novel 360deg semantic segmentation solution from a complementary perspective. Our network comprises three modules: a feature extraction module, a bi-directional compression module, and an ensemble decoding module. First, we extract multi-scale features from a panorama. Then, a bi-directional compression module is designed to compress features into two complementary low-dimensional representations, which provide content perception and distortion prior. Furthermore, to facilitate the fusion of bi-directional features, we design a unique self distillation strategy in the ensemble decoding module to enhance the interaction of different features and further improve the performance. Experimental results show that our approach outperforms the state-of-the-art solutions on quantitative evaluations while displaying the best performance on visual appearance.
**********
Out-of-Distribution Detection With Reconstruction Error and Typicality-Based Penalty
Genki Osada, Tsubasa Takahashi, Budrul Ahsan, Takashi Nishide; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5551-5563
The task of out-of-distribution (OOD) detection is vital to realize safe and reliable operation for real-world applications. After the failure of likelihood-based detection in high dimensions had been revealed, approaches based on the typical set have been attracting attention; however, they still have not achieved sat

isfactory performance. Beginning by presenting the failure case of the typicality-based approach, we propose a new reconstruction error-based approach that employs normalizing flow (NF). We further introduce a typicality-based penalty, and by incorporating it into the reconstruction error in NF, we propose a new OOD detection method, penalized reconstruction error (PRE). Because the PRE detects test inputs that lie off the in-distribution manifold, it also effectively detects adversarial examples. We show the effectiveness of our method through the evaluation using natural image datasets, CIFAR-10, TinyImageNet, and ILSVRC2012.
**********

Image-Consistent Detection of Road Anomalies As Unpredictable Patches
Tomáš Vojí■, Ji■í Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5491-5500
We propose a novel method for anomaly detection primarily aiming at autonomous driving. The design of the method, called DaCUP (Detection of anomalies as Consistent Unpredictable Patches), is based on two general properties of anomalous objects: an anomaly is (i) not from a class that could be modelled and (ii) it is not similar (in appearance) to non-anomalous objects in the image. To this end, we propose a novel embedding bottleneck in an auto-encoder like architecture that enables modelling of a diverse, multi-modal known class appearance (e.g. road). Secondly, we introduce novel image-conditioned distance features that allow known class identification in a nearest-neighbour manner on-the-fly, greatly increasing its ability to distinguish true and false positives. Lastly, an inpainting module is utilized to model the uniqueness of detected anomalies and significantly reduce false positives by filtering regions that are similar, thus reconstructable from their neighbourhood. We demonstrate that filtering of regions based on their similarity to neighbour regions, using e.g. an inpainting module, is general and can be used with other methods for reduction of false positives. The proposed method is evaluated on several publicly available datasets for road anomaly detection and on a maritime benchmark for obstacle avoidance. The method achieves state-of-the-art performance in both tasks with the same hyper-parameters with no domain specific design.
**********

Human-in-the-Loop Video Semantic Segmentation Auto-Annotation
Nan Qiao, Yuyin Sun, Chong Liu, Lu Xia, Jiajia Luo, Ke Zhang, Cheng-Hao Kuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5881-5891
Accurate per-pixel semantic class annotations of the entire video are crucial for designing and evaluating video semantic segmentation algorithms. However, the annotations are usually limited to a small subset of the video frames due to the high annotation cost and limited budget in practice. In this paper, we propose a novel human-in-the-loop framework called HVSA to generate semantic segmentation annotations for the entire video using only a small annotation budget. Our method alternates between active sample selection and test-time fine-tuning algorithms until annotation quality is satisfied. In particular, the active sample selection algorithm picks the most important samples to get manual annotations, where the sample can be a video frame, a rectangle, or even a super-pixel. Further, the test-time fine-tuning algorithm propagates the manual annotations of selected samples to the entire video. Real-world experiments show that our method generates highly accurate and consistent semantic segmentation annotations while simultaneously enjoys significantly small annotation cost.
**********

Deep Learning Methodology for Early Detection and Outbreak Prediction of Invasive Species Growth
Nathan Elias; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6335-6343
Invasive species (IS) cause major environmental damages, costing approximately 1.4 Trillion globally. Early detection and rapid response (EDRR) is key to mitigating IS growth, but current EDRR methods are highly inadequate at addressing IS growth. In this paper, a machine-learning-based approach to combat IS spread is proposed, in which identification, detection, and prediction of IS growth are au

tomated in a novel mobile application and scalable models. This paper details th e techniques used for the novel development of deep, multi-dimensional Convoluti onal Neural Networks (CNNs) to detect the presence of IS in both 2D and 3D space s, as well as the creation of geospatial Long Short-Term Memory (LSTMs) models t o then accurately quantify, simulate, and project invasive species' future envir onmental spread. Results from conducting training and in-field validation studie s show that this new methodology significantly improves current EDRR methods, by drastically decreasing the intensity of manual field labor while providing a to olkit that increases the efficiency and efficacy of ongoing efforts to combat IS . Furthermore, this research presents scalable expansion into dynamic LIDAR and aerial detection of IS growth, with the proposed toolkit already being deployed by state parks and national environmental/wildlife services.

**********

Contrastive Learning of Semantic Concepts for Open-Set Cross-Domain Retrieval

Aishwarya Agarwal, Srikrishna Karanam, Balaji Vasan Srinivasan, Biplab Banerjee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visio n (WACV), 2023, pp. 4115-4124

We consider the problem of image retrieval where query images during testing bel ong to classes and domains both unseen during training. This requires learning a feature space that has the ability to generalize across both classes and domain s. To this end, we propose semantic contrastive concept network (SCNNet), a new learning framework that helps take a step towards class and domain generalizatio n in a principled fashion. Unlike existing methods that rely on global object re presentations, SCNNet proposes to learn a set of local concept vectors to facili tate unseen-class generalization. To this end, SCNNet's key innovations include (a) a novel trainable local concept extraction module that learns an orthonormal set of basis vectors, and (b) computes local features for any unseen-class data as a linear combination of the learned basis set. Next, to enable unseen-domain generalization, SCNNet proposes to generate supervisory signals from an adjacen t data modality, i.e., natural language, by mining freely available textual labe l information associated with images. SCNNet derives these signals from our nove l trainable semantic ordinal distance constraints that ensure semantic consisten cy between pairs of images sampled from different domains. Both the proposed mod ules above enable end-to-end training of the SCNNet, resulting in a model that h elps establish state-of-the-art performance on the standard DomainNet, PACS, and Sketchy benchmark datasets with average Prec@200 improvements of 42.6%, 6.5%, a nd 13.6% respectively over the most recently reported results.

**********

Cross-Task Attention Mechanism for Dense Multi-Task Learning

Ivan Lopes, Tuan-Hung Vu, Raoul de Charette; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2329-2338

Multi-task learning has recently become a promising solution for a comprehensive understanding of complex scenes. With an appropriate design multi-task models c an not only be memory-efficient but also favour the exchange of complementary si gnals across tasks. In this work, we jointly address 2D semantic segmentation, a nd two geometry-related tasks, namely dense depth, surface normal estimation as well as edge estimation showing their benefit on indoor and outdoor datasets. We propose a novel multi-task learning architecture that exploits pair-wise cross-task exchange through correlation-guided attention and self-attention to enhance the average representation learning for all tasks. We conduct extensive experim ents considering three multi-task setups, showing the benefit of our proposal in comparison to competitive baselines in both synthetic and real benchmarks. We a lso extend our method to the novel multi-task unsupervised domain adaptation set ting. Our code is open-source.

**********

Online Adaptive Temporal Memory With Certainty Estimation for Human Trajectory P rediction

Manh Huynh, Gita Alaghband; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 940-949

Pedestrian trajectory prediction is an essential component of autonomous systems

and robot navigation. Recent research has shown promising predictive performanc
e by designing trajectory prediction networks to model a variety of motion-relat
ed features. Different from existing works, our focus is on designing a novel on
line adaptation framework (OATMem) to exploit the temporal similarities among tr
ajectory samples encountered during testing to improve the prediction accuracy o
f any such models (i.e., predictors) without knowing the details of these predic
tors. Our framework consists of two novel modules: an augmented temporal observe
-target memory network (ATM) and a certainty-based selector (CS). Inspired by th
e concept of key-value memory networks [16], we design the ATM to learn the temp
oral information from short-term past frames by encoding the trajectory samples
of past pedestrians in form of observe-target (i.e., key-value) during testing.
In addition, we propose a certainty-based selector (CS) to enhance the predictiv
e ability of our framework under scenarios where there are large temporal dissim
ilarities between current pedestrians' movements and those stored in memory. In
dynamic scenes, these scenarios commonly occur due to abrupt changes in contexts
, such as camera motions, scene contexts, and pedestrians' behaviors. We extensi
vely evaluate our framework in commonly-used datasets for pedestrian trajectory
prediction: JAAD [12] and PIE [19] and show that our framework significantly imp
roves the prediction accuracy of state-of-the-art models. Finally, in-depth abla
tion studies and analyses are conducted to show on the importance of each propos
ed component.
**********

FaceOff: A Video-to-Video Face Swapping System
Aditya Agarwal, Bipasha Sen, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C. V. J
awahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Compute
r Vision (WACV), 2023, pp. 3495-3504
Doubles play an indispensable role in the movie industry. They take the place of
the actors in dangerous stunt scenes or scenes where the same actor plays multi
ple characters. The double's face is later replaced with the actor's face and ex
pressions manually using expensive CGI technology, costing millions of dollars a
nd taking months to complete. An automated, inexpensive, and fast way can be to
use face-swapping techniques that aim to swap an identity from a source face vid
eo (or an image) to a target face video. However, such methods cannot preserve t
he source expressions of the actor important for the scene's context. To tackle
this challenge, we introduce video-to-video (V2V) face-swapping, a novel task of
face-swapping that can preserve (1) the identity and expressions of the source
(actor) face video and (2) the background and pose of the target (double) video.
We propose FaceOff, a V2V face-swapping system that operates by learning a robu
st blending operation to merge two face videos following the constraints above.
It reduces the videos to a quantized latent space and then blends them in the re
duced space. FaceOff is trained in a self-supervised manner and robustly tackles
the non-trivial challenges of V2V face-swapping. As shown in the experimental s
ection, FaceOff significantly outperforms alternate approaches qualitatively and
quantitatively.
**********

iColoriT: Towards Propagating Local Hints to the Right Region in Interactive Col
orization by Leveraging Vision Transformer
Jooyeol Yun, Sanghyeon Lee, Minho Park, Jaegul Choo; Proceedings of the IEEE/CVF
Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1787-179
6
Point-interactive image colorization aims to colorize grayscale images when a us
er provides the colors for specific locations. It is essential for point-interac
tive colorization methods to appropriately propagate user-provided colors (i.e.,
user hints) in the entire image to obtain a reasonably colorized image with min
imal user effort. However, existing approaches often produce partially colorized
results due to the inefficient design of stacking convolutional layers to propa
gate hints to distant relevant regions. To address this problem, we present iCol
oriT, a novel point-interactive colorization Vision Transformer capable of propa
gating user hints to relevant regions, leveraging the global receptive field of
Transformers. The self-attention mechanism of Transformers enables iColoriT to s

electively colorize relevant regions with only a few local hints. Our approach c olorizes images in real-time by utilizing pixel shuffling, an efficient upsampli ng technique that replaces the decoder architecture. Also, in order to mitigate the artifacts caused by pixel shuffling with large upsampling ratios, we present the local stabilizing layer. Extensive quantitative and qualitative results dem onstrate that our approach highly outperforms existing methods for point-interac tive colorization, producing accurately colorized images with a user's minimal e ffort. Official codes are available at https://pmh9960.github.io/research/iColor iT/.
**********

Cross-Modality Feature Fusion Network for Few-Shot 3D Point Cloud Classification
Minmin Yang, Jiajing Chen, Senem Velipasalar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 653-662
Recent years have witnessed significant progress in the field of few-shot image classification while few-shot 3D point cloud classification still remains under-explored. Real-world 3D point cloud data often suffers from occlusions, noise an d deformation, which make the few-shot 3D point cloud classification even more c hallenging. In this paper, we propose a cross-modality feature fusion network, f or few-shot 3D point cloud classification, which aims to recognize an object giv en only a few labeled samples, and provides better performance even with point c loud data with missing points. More specifically, we train two models in paralle l. One is a projection-based model with ResNet-18 as the backbone and the other one is a point-based model with a DGCNN backbone. Moreover, we design a Support-Query Mutual Attention (sqMA) module to fully exploit the correlation between su pport and query features. Extensive experiments on three datasets, namely ModelN et40, ModelNet40-C and ScanObjectNN, show the effectiveness of our method, and i ts robustness to missing points. Our proposed method outperforms different state -of-the-art baselines on all datasets. The margin of improvement is even larger on the ScanObjectNN dataset, which is collected from real-world scenes and is mo re challenging with objects having missing points.
**********

Training Auxiliary Prototypical Classifiers for Explainable Anomaly Detection in Medical Image Segmentation
Wonwoo Cho, Jeonghoon Park, Jaegul Choo; Proceedings of the IEEE/CVF Winter Conf erence on Applications of Computer Vision (WACV), 2023, pp. 2624-2633
Machine learning-based algorithms using fully convolutional networks (FCNs) have been a promising option for medical image segmentation. However, such deep netw orks silently fail if input samples are drawn far from the training data distrib ution, thus causing critical problems in automatic data processing pipelines. To overcome such out-of-distribution (OoD) problems, we propose a novel OoD score formulation and its regularization strategy by applying an auxiliary add-on clas sifier to an intermediate layer of an FCN, where the auxiliary module is helfpul for analyzing the encoder output features by taking their class information int o account. Our regularization strategy train the module along with the FCN via t he principle of outlier exposure so that our model can be trained to distinguish OoD samples from normal ones without modifying the original network architectur e. Our extensive experiment results demonstrate that the proposed approach can s uccessfully conduct effective OoD detection without loss of segmentation perform ance. In addition, our module can provide reasonable explanation maps along with OoD scores, which can enable users to analyze the reliability of predictions.
**********

Representation Recovering for Self-Supervised Pre-Training on Medical Images
Xiangyi Yan, Junayed Naushad, Shanlin Sun, Kun Han, Hao Tang, Deying Kong, Haoyu Ma, Chenyu You, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on A pplications of Computer Vision (WACV), 2023, pp. 2685-2695
Advances in self-supervised learning, especially in contrastive learning, have d rawn attention to investigating these techniques in providing effective visual r epresentations from unlabeled images. It enables the models' ability of extracti ng highly consistent features by generating different views. Due to the recent s uccess of Masked Autoencoders (MAE), an emerging trend of exploring generative m

odeling in self-supervised learning has come back into sight of the community. The generative approaches encode the input into a compact embedding and empower the models' ability of recovering the original input. However, in our experiments, we found vanilla MAE mainly recovers course high level semantic information and barely recovers detailed low level information. We show that in dense downstream prediction tasks like multi-organ segmentation, directly applying MAE is not ideal. In this paper, we propose RepRec, a hybrid visual representation learning framework for self-supervised pre-training on large-scale unlabelled medical datasets, which takes advantage of both contrastive and generative modeling. In our method, to solve the aforementioned dilemma that MAE encounters, a convolutional encoder is pre-trained to provide low-level feature information, in a contrastive way; and a transformer encoder is pre-trained to produce high level semantic dependency, in a generative way -- by recovering masked representations from the convolutional encoder. Extensive experiments on three multi-organ segmentation datasets demonstrate that our method outperforms current state-of-the-art methods.

**********

Physically Plausible Animation of Human Upper Body From a Single Image

Ziyuan Huang, Zhengping Zhou, Yung-Yu Chuang, Jiajun Wu, C. Karen Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 930-939

We present a new method for generating controllable, dynamically responsive, and photorealistic human animations. Given an image of a person, our system allows the user to generate Physically plausible Upper Body Animation (PUBA) using interaction in the image space, such as dragging their hand to various locations. We formulate a reinforcement learning problem to train a dynamic model that predicts the person's next 2D state (i.e., keypoints on the image) conditioned on a 3D action (i.e., joint torque), and a policy that outputs optimal actions to control the person to achieve desired goals. The dynamic model leverages the expressiveness of 3D simulation and the visual realism of 2D videos. PUBA generates 2D keypoint sequences that achieve task goals while being responsive to forceful perturbation. The sequences of keypoints are then translated by a pose-to-image generator to produce the final photorealistic video.

**********

D-Extract: Extracting Dimensional Attributes From Product Images

Pushpendu Ghosh, Nancy Wang, Promod Yenigalla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3641-3649

Product dimension is a crucial piece of information enabling customers make better buying decisions. E-commerce websites extract dimension attributes to enable customers filter the search results according to their requirements. The existing methods extract dimension attributes from textual data like title and product description. However, this textual information often exists in an ambiguous, disorganised structure. In comparison, images can be used to extract reliable and consistent dimensional information. With this motivation, we hereby propose two novel architecture to extract dimensional information from product images. The first namely Single-Box Classification Network is designed to classify each text token in the image, one at a time, whereas the second architecture namely Multi-Box Classification Network uses a transformer network to classify all the detected text tokens simultaneously. To attain better performance, the proposed architectures are also fused with statistical inferences derived from the product category which further increased the F1-score of the Single-Box Classification Network by  3.78% and Multi-Box Classification Network by  0.9%. We use distance supervision technique to create a large scale automated dataset for pretraining purpose and notice considerable improvement when the models were pretrained on the large data before finetuning. The proposed model achieves a desirable precision of 91.54% at 89.75% recall and outperforms the other state of the art approaches by  4.76% in F1-score.

**********

MASTAF: A Model-Agnostic Spatio-Temporal Attention Fusion Network for Few-Shot Video Classification

Xin Liu, Huanle Zhang, Hamed Pirsiavash, Xin Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2508-2517

We propose MASTAF, a Model-Agnostic Spatio-Temporal Attention Fusion network for few-shot video classification. MASTAF takes input from a general video spatial and temporal representation,e.g., using 2D CNN, 3D CNN, and Video Transformer. Then, to make the most of such representations, we use self- and cross-attention models to highlight the critical spatio-temporal region to increase the inter-class variations and decrease the intra-class variations. Last, MASTAF applies a lightweight fusion network and a nearest neighbor classifier to classify each query video. We demonstrate that MASTAF improves the state-of-the-art performance on three few-shot video classification benchmarks(UCF101, HMDB51, and Something-Something-V2), e.g., by up to 91.6%, 69.5%, and 60.7% for five-way one-shot video classification, respectively.

**********

Attribution-Aware Weight Transfer: A Warm-Start Initialization for Class-Incremental Semantic Segmentation

Dipam Goswami, René Schuster, Joost van de Weijer, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3195-3204

In class-incremental semantic segmentation (CISS), deep learning architectures suffer from the critical problems of catastrophic forgetting and semantic background shift. Although recent works focused on these issues, existing classifier initialization methods do not address the background shift problem and assign the same initialization weights to both background and new foreground class classifiers. We propose to address the background shift with a novel classifier initialization method which employs gradient-based attribution to identify the most relevant weights for new classes from the classifier's weights for the previous background and transfers these weights to the new classifier. This warm-start weight initialization provides a general solution applicable to several CISS methods. Furthermore, it accelerates learning of new classes while mitigating forgetting. Our experiments demonstrate significant improvement in mIoU compared to the state-of-the-art CISS methods on the Pascal-VOC 2012, ADE20K and Cityscapes datasets.

**********

Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID

Djebril Mekhazni, Maximilien Dufau, Christian Desrosiers, Marco Pedersoli, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1624-1633

Systems for person re-identification (ReID) can achieve a high level of accuracy when trained on large fully-labeled image datasets. However, the domain shift typically associated with diverse operational capture conditions (e.g., camera viewpoints and lighting) may translate to a significant decline in performance. This paper focuses on unsupervised domain adaptation (UDA) for video-based ReID -- a relevant scenario that is less explored in the literature. In this scenario, the ReID model must adapt to a complex target domain defined by a network of diverse video cameras based on tracklet information. State-of-art methods cluster unlabeled target data, yet domain shifts across target cameras (sub-domains) can lead to poor initialization of clustering methods that propagates noise across epochs, and the ReID model cannot accurately associate samples of the same identity. In this paper, an UDA method is introduced for video person ReID that leverages knowledge on video tracklets, and on the distribution of frames captured over target cameras to improve the performance of CNN backbones trained using pseudo-labels. Our method relies on an adversarial approach, where a camera-discriminator network is introduced to extract discriminant camera-independent representations, facilitating the subsequent clustering. In addition, a weighted contrastive loss is proposed to leverage the confidence of clusters, and mitigate the risk of incorrect identity associations. Experimental results obtained on three challenging video-based person ReID datasets -- PRID2011, iLIDS-VID, and MARS -- indicate that our proposed method can outperform related state-of-the-art methods.

The code is available at: https://github.com/wacv23775/775.
**********
Phantom Sponges: Exploiting Non-Maximum Suppression To Attack Deep Object Detectors
Avishag Shapira, Alon Zolfi, Luca Demetrio, Battista Biggio, Asaf Shabtai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4571-4580
Adversarial attacks against deep learning-based object detectors have been studied extensively in the past few years. Most of the attacks proposed have targeted the model's integrity (i.e., caused the model to make incorrect predictions), while adversarial attacks targeting the model's availability, a critical aspect in safety-critical domains such as autonomous driving, have not yet been explored by the machine learning research community. In this paper, we propose a novel attack that negatively affects the decision latency of an end-to-end object detection pipeline. We craft a universal adversarial perturbation (UAP) that targets a widely used technique integrated in many object detector pipelines - non-maximum suppression (NMS). Our experiments demonstrate the proposed UAP's ability to increase the processing time of individual frames by adding "phantom" objects that overload the NMS algorithm while preserving the detection of the original objects which allows the attack to go undetected for a longer period of time.
**********
Semi-Supervised Domain Adaptation With Auto-Encoder via Simultaneous Learning
Md Mahmudur Rahman, Rameswar Panda, Mohammad Arif Ul Alam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 402-411
We present a new semi-supervised domain adaptation framework that combines a novel auto-encoder-based domain adaptation model with a simultaneous learning scheme providing stable improvements over state-of-the-art domain adaptation models. Our framework holds strong distribution matching property by training both source and target auto-encoders using a novel simultaneous learning scheme on a single graph with an optimally modified MMD loss objective function. Additionally, we design a semi-supervised classification approach by transferring the aligned domain invariant feature spaces from source domain to the target domain. We evaluate on three datasets and show proof that our framework can effectively solve both fragile convergence (adversarial) and weak distribution matching problems between source and target feature space (discrepancy) with a high 'speed' of adaptation requiring a very low number of iterations.
**********
SAILOR: Scaling Anchors via Insights Into Latent Object Representation
Dušan Mali■, Christian Fruhwirth-Reisinger, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 623-632
LiDAR 3D object detection models are inevitably biased towards their training dataset. The detector clearly exhibits this bias when employed on a target dataset, particularly towards object sizes. However, object sizes vary heavily between domains due to, for instance, different labeling policies or geographical locations. State-of-the-art unsupervised domain adaptation approaches outsource methods to overcome the object size bias. Mainstream size adaptation approaches exploit target domain statistics, contradicting the original unsupervised assumption. Our novel unsupervised anchor calibration method addresses this limitation. Given a model trained on the source data, we estimate the optimal target anchors in a completely unsupervised manner. The main idea stems from an intuitive observation: by varying the anchor sizes for the target domain, we inevitably introduce noise or even remove valuable object cues. The latent object representation, perturbed by the anchor size, is closest to the learned source features only under the optimal target anchors. We leverage this observation for anchor size optimization. Our experimental results show that, without any retraining, we achieve competitive results even compared to state-of-the-art weakly-supervised size adaptation approaches. In addition, our anchor calibration can be combined with such existing methods, making them completely unsupervised.

**********
An Unified Framework for Language Guided Image Completion
Jihyun Kim, Seong-Hun Jeong, Kyeongbo Kong, Suk-Ju Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2568-2578
Image completion is a research field which aims to generate visual contents for unknown regions of an image. Image outpainting and wide-range image blending, which we refer to as extensive painting, are considered challenging because compared to the large unknown regions, relatively less context is provided. Some recent studies have tried to decrease the complexity of extensive painting by generating image hints for the missing regions. In this paper, we introduce a novel modality of hints, the natural language. Moreover, we propose a Captioning-based Extensive Painting (CEP) module, which combines models for two different multi-modal tasks: image captioning and text-guided image completion. In order to generate appropriate captions for masked images, the image captioning model is optimized using self-critical sequence training (SCST) method with random masks. The biggest benefit of our methodology is the accessibility to well-designed image captioning and text-guided image manipulation models such as OFA and GLIDE without the need for additional architectural changes. In evaluation, our model demonstrates remarkable performance even with complicated image datasets both quantitatively and qualitatively.
**********
Proactive Deepfake Defence via Identity Watermarking
Yuan Zhao, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, Xin Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4602-4611
The explosive progress of Deepfake techniques poses unprecedented privacy and security risks toward our society by creating real-looking but fake visual content. However, the current Deepfake detection studies are still in their infancy, because they mainly rely on capturing artifacts left by a Deepfake synthesis process as detection clues. These artifacts could be easily obscured due to various distortions (e.g. blurring) and could also be removed with the development of advanced Deepfake techniques, rendering the artifacts-based detection methods less effective in achieving reliable forgery forensics. In this paper, we propose a novel Deepfake detection method that does not depend on identifying the synthesized artifacts, but resorts to a mechanism of anti-counterfeit labels. Specifically, we design a neural network with an encoder-decoder structure to embed messages as anti-Deepfake labels into the facial identity features. Since the injected label is entangled with the facial identity feature, it will be sensitive to face swap translations (i.e., Deepfake), but robust to conventional image modifications (e.g., resize and compress). Therefore, we can check whether the watermarked image has been tampered with by Deepfake methods according to the existence of the label. Experimental results demonstrate that our method can achieve an average detection accuracy of more than 80%, which validates the effectiveness of the proposed method to implement Deepfake detection.
**********
CUDA-GHR: Controllable Unsupervised Domain Adaptation for Gaze and Head Redirection
Swati Jindal, Xin Eric Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 467-477
The robustness of gaze and head pose estimation models is highly dependent on the amount of labeled data. Recently, generative modeling has shown excellent results in generating photo-realistic images, which can alleviate the need for annotations. However, adopting such generative models to new domains while maintaining their ability to provide fine-grained control over different image attributes, e.g., gaze and head pose directions, has been a challenging problem. This paper proposes CUDA-GHR, an unsupervised domain adaptation framework that enables fine-grained control over gaze and head pose directions while preserving the appearance-related factors of the person. Our framework simultaneously learns to adapt to new domains and disentangle visual attributes such as appearance, gaze direc

tion, and head orientation by utilizing a label-rich source domain and an unlabe
led target domain. Extensive experiments on the benchmarking datasets show that
the proposed method can outperform state-of-the-art techniques on both quantitat
ive and qualitative evaluations. Furthermore, we demonstrate the effectiveness o
f generated image-label pairs in the target domain for pretraining networks for
the downstream task of gaze and head pose estimation. The source code and pre-tr
ained models are available at https://github.com/jswati31/cuda-ghr.
**********

Barlow Constrained Optimization for Visual Question Answering
Abhishek Jha, Badri Patro, Luc Van Gool, Tinne Tuytelaars; Proceedings of the IE
EE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 10
84-1093
Visual question answering is a vision-and-language multimodal task, that aims at
 predicting answers given samples from the question and image modalities. Most r
ecentmethods focus on learning a good joint embedding space ofimages and questio
ns, either by improving the interactionbetween these two modalities, or by makin
g it a more discriminant space. However, how informative this joint space is, ha
s not been well explored. In this paper, we propose a novel regularization for V
QA models, Constrained Optimization using Barlow's theory (COB), that improves t
he information content of the joint space by minimizing the redundancy. It reduc
es the correlation between the learned feature components and thereby disentangl
es semantic concepts. Our model also aligns the joint space with the answer embe
dding space, where we consider the answer and image+question as two different 'v
iews' of what in essence is the same semantic information. We propose a constrai
ned optimization policy to balance the categorical and redundancy minimization f
orces. When built on the state-of-the-art GGE model, the resulting model improve
s VQA accuracy by 1.4% and 4% on the VQA-CP v2 and VQA v2 datasets respectively.
 The model also exhibits better interpretability. Code is made available: https:
//github.com/abskjha/Barlow-constrained-VQA
**********

Few-Shot Learning of Compact Models via Task-Specific Meta Distillation
Yong Wu, Shekhor Chanda, Mehrdad Hosseinzadeh, Zhi Liu, Yang Wang; Proceedings o
f the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023
, pp. 6265-6274
We consider a new problem of few-shot learning of compact models. Meta-learning
 is a popular approach for few-shot learning. Previous work in meta-learning typi
cally assumes that the model architecture during meta-training is the same as th
e model architecture used for final deployment. In this paper, we challenge this
 basic assumption. For final deployment, we often need the model to be small. Bu
t small models usually do not have enough capacity to effectively adapt to new t
asks. In the mean time, we often have access to the large dataset and extensive
computing power during meta-training since meta-training is typically performed
on a server. In this paper, we propose task-specific meta distillation that simu
ltaneously learns two models in meta-learning: a large teacher model and a small
 student model. These two models are jointly learned during meta-training. Given
 a new task during meta-testing, the teacher model is first adapted to this task
, then the adapted teacher model is used to guide the adaptation of the student
model. The adapted student model is used for final deployment. We demonstrate th
e effectiveness of our approach in few-shot image classification using model-agn
ostic meta-learning (MAML). Our proposed method outperforms other alternatives o
n several benchmark datasets.
**********

OCR-VQGAN: Taming Text-Within-Image Generation
Juan A. Rodríguez, David Vazquez, Issam Laradji, Marco Pedersoli, Pau Rodriguez;
 Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visio
n (WACV), 2023, pp. 3689-3698
Synthetic image generation has recently experienced significant improvements in
domains such as natural image or art generation. However, the problem of figure
and diagram generation remains unexplored. A challenging aspect of generating fi
gures and diagrams is effectively rendering readable texts within the images. To

alleviate this problem, we present OCR-VQGAN, an image encoder, and decoder tha t leverages OCR pre-trained features to optimize a text perceptual loss, encoura ging the architecture to preserve high-fidelity text and diagram structure. To e xplore our approach, we introduce the Paper2Fig100k dataset, with over 100k imag es of figures and texts from research papers. The figures show architecture diag rams and methodologies of articles available at arXiv.org from fields like artif icial intelligence and computer vision. Figures usually include text and discret e objects, e.g., boxes in a diagram, with lines and arrows that connect them. We demonstrate the effectiveness of OCR-VQGAN by conducting several experiments on the task of figure reconstruction. Additionally, we explore the qualitative and quantitative impact of weighting different perceptual metrics in the overall lo ss function. We release code, models, and dataset at https://github.com/joanrod/ ocr-vqgan.

**********

PINER: Prior-Informed Implicit Neural Representation Learning for Test-Time Adap tation in Sparse-View CT Reconstruction

Bowen Song, Liyue Shen, Lei Xing; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1928-1938

Recently, deep learning has been introduced to solve important medical image rec onstruction problems such as sparse-view CT reconstruction. However, the develop ed deep reconstruction models are generally limited in generalization when appli ed to unseen testing samples in target domain. Furthermore, privacy concerns may impede the availability of source-domain training data to retrain or adapt the model to the target-domain testing data, which are quite common in real-world me dical applications. To address these issues, we introduce a source-free black-bo x test-time adaptation method for sparse-view CT reconstruction with unknown noi se levels based on prior-informed implicit neural representation learning (PINER ). By leveraging implicit neural representation learning to generate the image r epresentations at various noise levels, the proposed method is able to construct the adapted input representations at test time based on the inference of black- box model and output analysis. We performed experiments of source-free test-time adaptation for sparse-view CT reconstruction with unknown noise levels on multi ple anatomical sites with different black-box deep reconstruction models, where our method outperforms the state-of-the-art algorithms.

**********

DE-CROP: Data-Efficient Certified Robustness for Pretrained Classifiers

Gaurav Kumar Nayak, Ruchit Rawal, Anirban Chakraborty; Proceedings of the IEEE/C VF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4622-4 631

Certified defense using randomized smoothing is a popular technique to provide r obustness guarantees for deep neural networks against l2 adversarial attacks. Ex isting works use this technique to provably secure a pretrained non-robust model by training a custom denoiser network on entire training data. However, access to the training set may be restricted to a handful of data samples due to constr aints such as high transmission cost and the proprietary nature of the data. Thu s, we formulate a novel problem of "how to certify the robustness of pretrained models using only a few training samples". We observe that training the custom d enoiser directly using the existing techniques on limited samples yields poor ce rtification. To overcome this, our proposed approach (DE-CROP) generates class-b oundary and interpolated samples corresponding to each training sample, ensuring high diversity in the feature space of the pretrained classifier. We train the denoiser by maximizing the similarity between the denoised output of the generat ed sample and the original training sample in the classifier's logit space. We a lso perform distribution level matching using domain discriminator and maximum m ean discrepancy that yields further benefit. In white box setup, we obtain signi ficant improvements over the baseline on multiple benchmark datasets and also re port similar performance under the challenging black box setup.

**********

Few-Shot Object Counting With Similarity-Aware Feature Enhancement

Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, Xinyi Le; Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6315-6324

This work studies the problem of few-shot object counting, which counts the number of exemplar objects (i.e., described by one or several support images) occurring in the query image. The major challenge lies in that the target objects can be densely packed in the query image, making it hard to recognize every single one. To tackle the obstacle, we propose a novel learning block, equipped with a similarity comparison module and a feature enhancement module. Concretely, given a support image and a query image, we first derive a score map by comparing their projected features at every spatial position. The score maps regarding all support images are collected together and normalized across both the exemplar dimension and the spatial dimensions, producing a reliable similarity map. We then enhance the query feature with the support features by employing the developed point-wise similarities as the weighting coefficients. Such a design encourages the model to inspect the query image by focusing more on the regions akin to the support images, leading to much clearer boundaries between different objects. Extensive experiments on various benchmarks and training setups suggest that we surpass the state-of-the-art methods by a sufficiently large margin. For instance, on a recent large-scale FSC-147 dataset, we surpass the state-of-the-art method by improving the mean absolute error from 22.08 to 14.32. Code has been released in https://github.com/zhiyuanyou/SAFECount.

**********

PSENet: Progressive Self-Enhancement Network for Unsupervised Extreme-Light Image Enhancement

Hue Nguyen, Diep Tran, Khoi Nguyen, Rang Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1756-1765

The extremes of lighting (e.g. too much or too little light) usually cause many troubles for machine and human vision. Many recent works have mainly focused on under-exposure cases where images are often captured in low-light conditions (e.g. nighttime) and achieved promising results for enhancing the quality of images. However, they are inferior to handling images under over-exposure. To mitigate this limitation, we propose a novel unsupervised enhancement framework which is robust against various lighting conditions while does not require any well-exposed images to serve as the ground-truths. Our main concept is to construct pseudo-ground-truth images synthesized from multiple source images that simulate all potential exposure scenarios to train the enhancement network. Our extensive experiments show that the proposed approach consistently outperforms the current state-of-the-art unsupervised counterparts in several public datasets in terms of both quantitative metrics and qualitative results. Our code is available at https://github.com/VinAIResearch/PSENet-Image-Enhancement

**********

BrightFlow: Brightness-Change-Aware Unsupervised Learning of Optical Flow

Rémi Marsal, Florian Chabot, Angélique Loesch, Hichem Sahbi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2061-2070

Unsupervised optical flow estimation relies on the assumption that pixels characterizing the same observed object should exhibit a stable appearance across video frames. With this assumption, the long-standing principle behind flow estimation consists in optimizing a photometric loss that maximizes the similarity between paired pixels in successive frames. However, these frames could be subject to strong brightness changes due to the radiometric properties of scenes as well as their viewing conditions. In this paper, we present BrightFlow, a new method to train any optical flow estimation network in an unsupervised manner. It consists in training two networks that jointly estimate optical flow and brightness changes. These changes are then compensated in the photometric loss so that reconstruction errors due to shadows or reflections will not affect negatively the training. As this compensation mechanism is only used at training stage, our method does not impact the number of parameters or the complexity at inference. Extensive experiments conducted on standard datasets and optical flow architectures show a consistent gain of our method. Source code is available at https://github.c

om/CEA-LIST/BrightFlow.
**********
Event-Specific Audio-Visual Fusion Layers: A Simple and New Perspective on Video Understanding
Arda Senocak, Junsik Kim, Tae-Hyun Oh, Dingzeyu Li, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2237-2247
To understand our surrounding world, our brain is continuously inundated with multisensory information and their complex interactions coming from the outside world at any given moment. While processing this information might seem effortless for human brains, it is challenging to build a machine that can perform similar tasks since complex interactions cannot be dealt with a single type of integration but require more sophisticated approaches. In this paper, we propose a new simple method to address the multisensory integration in video understanding. Unlike previous works where a single fusion type is used, we design a multi-head model with individual event-specific layers to deal with different audio-visual relationships, enabling different ways of audio-visual fusion. Experimental results show that our event-specific layers can discover unique properties of the audio-visual relationships in the videos, e.g., semantically matched moments, and rhythmic events. Moreover, although our network is trained with single labels, our multi-head design can inherently output additional semantically meaningful multi-labels for a video. As an application, we demonstrate that our proposed method can expose the extent of event-characteristics of popular benchmark datasets.
**********
Transformers for Recognition in Overhead Imagery: A Reality Check
Francesco Luzi, Aneesh Gupta, Leslie Collins, Kyle Bradbury, Jordan Malof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3778-3787
There is evidence that transformers offer state-of-the-art recognition performance on tasks involving overhead imagery (e.g., satellite imagery). However, it is difficult to make unbiased empirical comparisons between competing deep learning models, making it unclear whether, and to what extent, transformer-based models are beneficial. In this paper we systematically compare the impact of adding transformer structures into state-of-the-art segmentation models for overhead imagery. Each model is given a similar budget of free parameters, and their hyperparameters are optimized using Bayesian Optimization with a fixed quantity of data and computation time. We conduct our experiments with a large and diverse dataset comprising two large public benchmarks: Inria and DeepGlobe. We perform additional ablation studies to explore the impact of specific transformer-based modeling choices. Our results suggest that transformers provide consistent, but modest, performance improvements. We only observe this advantage however in hybrid models that combine convolutional and transformer-based structures, while fully transformer-based models achieve relatively poor performance.
**********
Bent & Broken Bicycles: Leveraging Synthetic Data for Damaged Object Re-Identification
Luca Piano, Filippo Gabriele Pratticò, Alessandro Sebastian Russo, Lorenzo Lanari, Lia Morra, Fabrizio Lamberti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4881-4891
Instance-level object re-identification is a fundamental computer vision task, with applications from image retrieval to intelligent monitoring and fraud detection. In this work, we propose the novel task of damaged object re-identification, which aims at distinguishing changes in visual appearance due to deformations or missing parts from subtle intra-class variations. To explore this task, we leverage the power of computer-generated imagery to create, in a semi-automatic fashion, high-quality synthetic images of the same bike before and after a damage occurs. The resulting dataset, Bent & Broken Bicycles (BBBicycles), contains 39,200 images and 2,800 unique bike instances spanning 20 different bike models. As a baseline for this task, we propose TransReI3D, a multi-task, transformer-based deep network unifying damage detection (framed as a multi-label classification

task) with object re-identification.
**********
Efficient Flow-Guided Multi-Frame De-Fencing
Stavros Tsogkas, Fengjia Zhang, Allan Jepson, Alex Levinshtein; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1838-1847
Taking photographs "in-the-wild" is often hindered by fence obstructions that stand between the camera user and the scene of interest, and which are hard or impossible to avoid. De-fencing is the algorithmic process of automatically removing such obstructions from images, revealing the invisible parts of the scene. While this problem can be formulated as a combination of fence segmentation and image inpainting, this often leads to implausible hallucinations of the occluded regions. Existing multi-frame approaches rely on propagating information to a selected keyframe from its temporal neighbors, but they are often inefficient and struggle with alignment of severely obstructed images. In this work we draw inspiration from the video completion literature and develop a simplified framework for multi-frame de-fencing that computes high quality flow maps directly from obstructed frames and uses them to accurately align frames. Our primary focus is efficiency and practicality in a real-world setting: the input to our algorithm is a short image burst (5 frames) -- a data modality commonly available in modern smartphones-- and the output is a single reconstructed keyframe, with the fence removed. Our approach leverages simple yet effective CNN modules, trained on carefully generated synthetic data, and outperforms more complicated alternatives real bursts, both quantitatively and qualitatively, while running real-time.
**********
Adaptively-Realistic Image Generation From Stroke and Sketch With Diffusion Model
Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, Hsin-Ying Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4054-4062
Generating images from hand-drawings is a crucial and fundamental task in content creation. The translation is difficult as there exist infinite possibilities and the different users usually expect different outcomes. Therefore, we propose a unified framework supporting a three-dimensional control over the image synthesis from sketches and strokes based on diffusion models. Users can not only decide the level of faithfulness to the input strokes and sketches, but also the degree of realism, as the user inputs are usually not consistent with the real images. Qualitative and quantitative experiments demonstrate that our framework achieves state-of-the-art performance while providing flexibility in generating customized images with control over shape, color, and realism. Moreover, our method unleashes applications such as editing on real images, generation with partial sketches and strokes, and multi-domain multi-modal synthesis.
**********
Unifying Margin-Based Softmax Losses in Face Recognition
Yang Zhang, Simao Herdade, Kapil Thadani, Eric Dodds, Jack Culpepper, Yueh-Ning Ku; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3548-3557
In this work, we develop a theoretical and experimental framework to study the effect of margin penalties on angular softmax losses, which have led to state-of-the-art performance in face recognition. We also introduce a new multiplicative margin which performs comparably to previously proposed additive margins when the model is trained to convergence. A regime of the margin parameters can lead to degenerate minima, but these can be reliably avoided through the use of two regularization techniques that we propose. Our theory predicts the minimal angular distance between sample embeddings and the correct and wrong class prototype vectors learned during training, and it suggests a new method to identify optimal margin parameters without expensive tuning. Finally, we conduct a thorough ablation study of the margin parameters in our proposed framework, and we characterize the sensitivity of generalization to each parameter both theoretically and through experiments on standard face recognition benchmarks.

**********
Searching for Robust Binary Neural Networks via Bimodal Parameter Perturbation
Daehyun Ahn, Hyungjun Kim, Taesu Kim, Eunhyeok Park, Jae-Joon Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2410-2419
Binary neural networks (BNNs) are advantageous in performance and memory footprint but suffer from low accuracy due to their limited expression capability. Recent works have tried to enhance the accuracy of BNNs via a gradient-based search algorithm and showed promising results. However, the mixture of architecture search and binarization induce the instability of the search process, resulting in convergence to the suboptimal point. To address this issue, we propose a BNN architecture search framework with bimodal parameter perturbation. The bimodal parameter perturbation can improve the stability of gradient-based architecture search by reducing the sharpness of the loss surface along both weight and architecture parameter axes. In addition, we refine the inverted bottleneck convolution block for having robustness with BNNs. The synergy of the refined space and the stabilized search process allows us to find out the accurate BNNs with high computation efficiency. Experimental results show that our framework finds the best architecture on CIFAR-100 and ImageNet datasets in the existing search space for BNNs. We also tested our framework on another search space based on the inverted bottleneck convolution block, and the selected BNN models using our approach achieved the highest accuracy on both datasets with a much smaller number of equivalent operations than previous works.
**********
GaIA: Graphical Information Gain Based Attention Network for Weakly Supervised Point Cloud Semantic Segmentation
Min Seok Lee, Seok Woo Yang, Sung Won Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 582-591
While point cloud semantic segmentation is a significant task in 3D scene understanding, this task demands a time-consuming process of fully annotating labels. To address this problem, recent studies adopt a weakly supervised learning approach under the sparse annotation. Different from the existing studies, this study aims to reduce the epistemic uncertainty measured by the entropy for a precise semantic segmentation. We propose the graphical information gain based attention network called GaIA, which alleviates the entropy of each point based on the reliable information. The graphical information gain discriminates the reliable point by employing relative entropy between target point and its neighborhoods. We further introduce anchor-based additive angular margin loss, ArcPoint. The ArcPoint optimizes the unlabeled points containing high entropy towards semantically similar classes of the labeled points on hypersphere space. Experimental results on S3DIS and ScanNet-v2 datasets demonstrate our framework outperforms the existing weakly supervised methods.
**********
Unsupervised Multi-Object Segmentation Using Attention and Soft-Argmax
Bruno Sauvalle, Arnaud de La Fortelle; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3267-3276
We introduce a new architecture for unsupervised object-centric representation learning and multi-object detection and segmentation, which uses a translation-equivariant attention mechanism to predict the coordinates of the objects present in the scene and to associate a feature vector to each object. A transformer encoder handles occlusions and redundant detections, and a convolutional autoencoder is in charge of background reconstruction. We show that this architecture significantly outperforms the state of the art on complex synthetic benchmarks.
**********
Towards a Framework for Privacy-Preserving Pedestrian Analysis
Anil Kunchala, Mélanie Bouroche, Bianca Schoen-Phelan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4370-4380
The design of pedestrian-friendly infrastructures plays a crucial role in creating sustainable transportation in urban environments. Analyzing pedestrian behavi

our in response to existing infrastructure is pivotal to planning, maintaining, and creating more pedestrian-friendly facilities. Many approaches have been proposed to extract such behaviour by applying deep learning models to video data. Video data, however, includes an broad spectrum of privacy-sensitive information about individuals, such as their location at a given time or who they are with. Most of the existing models use privacy-invasive methodologies to track, detect, and analyse individual or group pedestrian behaviour patterns. As a step towards privacy-preserving pedestrian analysis, this paper introduces a framework to anonymize all pedestrians before analyzing their behaviors. The proposed framework leverages recent developments in 3D wireframe reconstruction and digital in-painting to represent pedestrians with quantitative wireframes by removing their images while preserving pose, shape, and background scene context. To evaluate the proposed framework, a generic metric is introduced for each of privacy and utility. Experimental evaluation on widely-used datasets shows that the proposed framework outperforms traditional and state-of-the-art image filtering approaches by generating best privacy utility trade-off.

**********

SD-Conv: Towards the Parameter-Efficiency of Dynamic Convolution

Shwai He, Chenbo Jiang, Daize Dong, Liang Ding; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6454-6463

Dynamic convolution achieves better performance for efficient CNNs at the cost of negligible FLOPs increase. However, the performance increase can not match the significantly expanded number of parameters, which is the main bottleneck in real-world applications. Contrastively, mask-based unstructured pruning obtains a lightweight network by removing redundancy in the heavy network. In this paper, we propose a new framework, Sparse Dynamic Convolution (SD-Conv), to naturally integrate these two paths such that it can inherit the advantage of dynamic mechanism and sparsity. We first design a binary mask derived from a learnable threshold to prune static kernels, significantly reducing the parameters and computational cost but achieving higher performance in Imagenet-1K. We further transfer pretrained models into a variety of downstream tasks, showing consistently better results than baselines. We hope our SD-Conv could be an efficient alternative to conventional dynamic convolutions.

**********

No Reference Opinion Unaware Quality Assessment of Authentically Distorted Images

Nithin C. Babu, Vignesh Kannan, Rajiv Soundararajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2459-2468

The quality assessment (QA) of camera captured authentically distorted images is important on account of its ubiquitous applications and challenging due to the lack of a reference. While there exists a plethora of supervised no reference (NR) image QA (IQA) algorithms, there is a need to study unsupervised or opinion unaware algorithms on account of their superior generalization performance. We explore self-supervised learning (SSL) for the feature design on authentically distorted images to predict quality without training on human labels. While SSL on synthetic distortions has recently shown promise, there is a need to enrich the feature learning on authentic distortions. The key challenge in achieving this is in the learning of quality sensitive features with mitigated content dependence. We design a self-supervised contrastive learning approach which only requires positives and introduce a content separation loss by estimating a bound on the mutual information between the features learnt and the content information. We show on multiple authentically distorted datasets that our self-supervised features can predict image quality by comparing with a corpus of pristine images and achieve state-of-the-art performance.

**********

3D-SpLineNet: 3D Traffic Line Detection Using Parametric Spline Representations

Maximilian Pittner, Alexandru Condurache, Joel Janai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 602-611

Monocular 3D traffic line detection jointly tackles the detection of lane markin

gs and regression of their 3D location. The greatest challenge is the exact estimation of various line shapes in the world, which highly depends on the chosen representation. While anchor-based and grid-based line representations have been proposed, all suffer from the same limitation, the necessity of discretizing the 3D space. To address this limitation, we present an anchor-free parametric lane representation, which defines traffic lines as continuous curves in 3D space. Choosing splines as our representation, we show their superiority over polynomials of different degrees that were proposed in previous 2D lane detection approaches. Our continuous representation allows us to model even complex lane shapes at any position in the 3D space, while implicitly enforcing smoothness constraints. Our model is validated on a synthetic 3D lane dataset including a variety of scenes in terms of complexity of road shape and illumination. We outperform the state-of-the-art in nearly all geometric performance metrics and achieve a great leap in the detection rate. In contrast to discrete representations, our parametric model requires no post-processing achieving highest processing speed. Additionally, we provide a thorough analysis over different parametric representations for 3D lane detection. The code and trained models are available on our project website https://3d-splinenet.github.io/.

**********

EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement

Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, Daniel McDuff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5008-5017

Camera-based physiological measurement is a growing field with neural models providing state-of-the-art performance. Prior research has explored various end-to-end architectures; however these methods still require several preprocessing steps and are not able to run directly on mobile and edge devices. The operations are often non-trivial to implement, making replication and deployment difficult and can even have a higher computational budget than the core network itself. In this paper, we propose two novel and efficient neural models for camera-based physiological measurement called EfficientPhys that remove the need for face detection, segmentation, normalization, color space transformation or any other preprocessing steps. Using an input of raw video frames, our models achieve strong accuracy on three public datasets. We show that this is the case whether using a transformer or convolutional backbone. We further evaluate the latency of the proposed networks and show that our most lightweight network also achieves a 33% improvement in efficiency.

**********

Arbitrary Style Guidance for Enhanced Diffusion-Based Text-to-Image Generation

Zhihong Pan, Xin Zhou, Hao Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4461-4471

Diffusion-based text-to-image generation models like GLIDE and DALLE-2 have gained wide success recently for their superior performance in turning complex text inputs into images of high quality and wide diversity. In particular, they are proven to be very powerful in creating graphic arts of various formats and styles. Although current models supported specifying style formats like oil painting or pencil drawing, fine-grained style features like color distributions and brush strokes are hard to specify as they are randomly picked from a conditional distribution based on the given text input. Here we propose a novel style guidance method to support generating images using arbitrary style guided by a reference image. The generation method does not require a separate style transfer model to generate desired styles while maintaining image quality in generated content as controlled by the text input. Additionally, the guidance method can be applied without a style reference, denoted as self style guidance, to generate images of more diverse styles. Comprehensive experiments prove that the proposed method remains robust and effective in a wide range of conditions, including diverse graphic art forms, image content types and diffusion models.

**********

Pik-Fix: Restoring and Colorizing Old Photos

Runsheng Xu, Zhengzhong Tu, Yuanqi Du, Xiaoyu Dong, Jinlong Li, Zibo Meng, Jiaqi Ma, Alan Bovik, Hongkai Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1724-1734
Restoring and inpainting the visual memories that are present, but often impaired, in old photos remains an intriguing but unsolved research topic. Decades-old photos often suffer from severe and commingled degradation such as cracks, defocus, and color-fading, which are difficult to treat individually and harder to repair when they interact. Deep learning presents a plausible avenue, but the lack of large-scale datasets of old photos makes addressing this restoration task very challenging. Here we present a novel reference-based end-to-end learning framework that is able to both repair and colorize old, degraded pictures. Our proposed framework consists of three modules: a restoration sub-network that conducts restoration from degradations, a similarity network that performs color histogram matching and color transfer, and a colorization subnet that learns to predict the chroma elements of images conditioned on chromatic reference signals. The overall system makes uses of color histogram priors from reference images, which greatly reduces the need for large-scale training data. We have also created a first-of-a-kind public dataset of real old photos that are paired with ground truth "pristine" photos that have been manually restored by PhotoShop experts. We conducted extensive experiments on this dataset and synthetic datasets, and found that our method significantly outperforms previous state-of-the-art models using both qualitative comparisons and quantitative measurements. The code is available at https://github.com/DerrickXuNu/Pik-Fix.
**********

FFM: Injecting Out-of-Domain Knowledge via Factorized Frequency Modification
Zijian Wang, Yadan Luo, Zi Huang, Mahsa Baktashmotlagh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4135-4144
This work addresses the Single Domain Generalization (SDG) problem, and aims to generalize a model from a single source (i.e., training) domain to multiple target (i.e., test) domains with different distributions. Most of the existing SDG approaches aim at generating out-of-domain samples by either transforming the source images into different styles or optimizing adversarial noise perturbations. In this paper, we show that generating images with diverse styles can be complementary to creating hard samples when tackling the SDG task. This inspires us to propose our approach of Factorized Frequency Modification (FFM) which can fulfil the requirement of generating diverse and hard samples to tackle the problem of out-of-domain generalization. Specifically, we design a unified framework consisting of a style transformation module, an adversarial perturbation module, and a dynamic frequency selection module. We seamlessly equip the framework with iterative adversarial training which facilitates the task model to learn discriminative features from hard and diverse augmented samples. We perform extensive experiments on four image recognition benchmark datasets of Digits-DG, CIFAR-10-C, CIFAR-100-C, and PACS, which demonstrates that our method outperforms existing state-of-the-art approaches.
**********

PatchZero: Defending Against Adversarial Patch Attacks by Detecting and Zeroing the Patch
Ke Xu, Yao Xiao, Zhaoheng Zheng, Kaijie Cai, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4632-4641
Adversarial patch attacks mislead neural networks by injecting adversarial pixels within a local region. Patch attacks can be highly effective in a variety of tasks and physically realizable via attachment (e.g. a sticker) to the real-world objects. Despite the diversity in attack patterns, adversarial patches tend to be highly textured and different in appearance from natural images. We exploit this property and present PatchZero, a general defense pipeline against white-box adversarial patches without retraining the downstream classifier or detector. Specifically, our defense detects adversaries at the pixel-level and "zeros out" the patch region by repainting with mean pixel values. We further design a two-s

tage adversarial training scheme to defend against the stronger adaptive attacks . PatchZero achieves SOTA defense performance on the image classification (Image Net, RESISC45), object detection (PASCAL VOC), and video classification (UCF101) tasks with little degradation in benign performance. In addition, PatchZero tra nsfers to different patch shapes and attack types.

**********

Multi-Scale Cell-Based Layout Representation for Document Understanding

Yuzhi Shi, Mijung Kim, Yeongnam Chae; Proceedings of the IEEE/CVF Winter Confere nce on Applications of Computer Vision (WACV), 2023, pp. 3670-3679

Deep learning techniques have achieved remarkable progress in document understan ding. Most models use coordinates to represent absolute or relative spatial info rmation of components, but they are difficult to represent latent rules in the d ocument layout. This makes learning layout representation to be more difficult. Unlike the previous researches which have employed the coordinate system, graph or grid to represent the document layout, we propose a novel layout representati on, the cell-based layout, to provide easy-to-understand spatial information for backbone models. In line with human reading habits, it uses cell information, i .e. row and column index, to represent the position of components in a document, and makes the document layout easier to understand. Furthermore, we proposed th e multi-scale layout to represent the hierarchical structure of layout, and deve loped a data augmentation method to improve the performance. Experiment results show that our method achieves the state-of-the-art performance in text-based tas ks, including form understanding and receipt understanding, and improves the per formance in image-based task such as document image classification. We released the code in the repo.

**********

Compact and Optimal Deep Learning With Recurrent Parameter Generators

Jiayun Wang, Yubei Chen, Stella X. Yu, Brian Cheung, Yann LeCun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3900-3910

Deep learning has achieved tremendous success by training increasingly large mod els, which are then compressed for practical deployment. We propose a drasticall y different approach to compact and optimal deep learning: We decouple the Degre es of freedom (DoF) and the actual number of parameters of a model, optimize a s mall DoF with predefined random linear constraints for a large model of an arbit rary architecture, in one-stage end-to-end learning. Specifically, we create a r ecurrent parameter generator (RPG), which repeatedly fetches parameters from a r ing and unpacks them onto a large model with random permutation and sign flippin g to promote parameter decorrelation. We show that gradient descent can automati cally find the best model under constraints with in fact faster convergence. Our extensive experimentation reveals a log-linear relationship between model DoF a nd accuracy. Our RPG demonstrates remarkable DoF reduction, and can be further p runed and quantized for additional run-time performance gain. For example, in te rms of top-1 accuracy on ImageNet, RPG achieves 96% of ResNet18's performance wi th only 18% DoF (the equivalent of one convolutional layer) and 52% of ResNet34' s performance with only 0.25% DoF! Our work shows significant potential of const rained neural optimization in compact and optimal deep learning.

**********

Simultaneous Acquisition of High Quality RGB Image and Polarization Information Using a Sparse Polarization Sensor

Teppei Kurita, Yuhi Kondo, Legong Sun, Yusuke Moriuchi; Proceedings of the IEEE/ CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 178-1 88

This paper proposes a novel polarization sensor structure and network architectu re to obtain a high-quality RGB image and polarization information. Conventional polarization sensors can simultaneously acquire RGB images and polarization inf ormation, but the polarizers on the sensor degrade the quality of the RGB images . There is a trade-off between the quality of the RGB image and polarization inf ormation as fewer polarization pixels reduce the degradation of the RGB image bu t decrease the resolution of polarization information. Therefore, we propose an

approach that resolves the trade-off by sparsely arranging polarization pixels o n the sensor and compensating for low-resolution polarization information with h igher resolution using the RGB image as a guide. Our proposed network architectu re consists of an RGB image refinement network and a polarization information co mpensation network. We confirmed the superiority of our proposed network in comp ensating the differential component of polarization intensity by comparing its p erformance with state-of-the-art methods for similar tasks: depth completion. Fu rthermore, we confirmed that our approach could simultaneously acquire higher qu ality RGB images and polarization information than conventional polarization sen sors, resolving the trade-off between the quality of RGB images and polarization information. The baseline code and newly generated real and synthetic large-sca le polarization image datasets are available for further research and developmen t.

**********

Skew-Robust Human-Object Interactions in Videos

Apoorva Agarwal, Rishabh Dabral, Arjun Jain, Ganesh Ramakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5098-5107

Humans are, arguably, one of the most important regions of interest in a visual analysis pipeline. Detecting how the human interacts with the surrounding enviro nment, thus, becomes an important problem and has several potential use-cases. W hile this has been adequately addressed in the literature in the image setting, there exist very few methods addressing the case for in-the-wild videos. The pro blem is further exacerbated by the high degree of label skew. To this end, we pr opose SeRVo-HOI, a robust end-to-end framework for recognizing human-object inte ractions from a video, particularly in high label-skew settings. The network con textualizes multiple image representations and is trained to explicitly handle d ataset skew. We propose and analyse methods to address the long-tail distributio n of the labels and show improvements on the tail-labels. SeRVo-HOI outperforms the state-of-the-art by a significant margin 21.1% vs 17.6% mAP on the large-sca le, in-the-wild VidHOI dataset while particularly demonstrating solid improvemen ts in the tail-classes 19.9% vs 17.3% mAP.

**********

Video Joint Denoising and Demosaicing With Recurrent CNNs

Valéry Dewil, Adrien Courtois, Mariano Rodríguez, Thibaud Ehret, Nicola Brandoni sio, Denis Bujoreanu, Gabriele Facciolo, Pablo Arias; Proceedings of the IEEE/CV F Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5108-51 19

Denoising and demosaicing are two critical components of the image/video process ing pipeline. While historically these two tasks have mainly been considered sep arately, current neural network approaches allow to obtain state-of-the-art resu lts by treating them jointly. However, most existing research focuses in single image or burst joint denoising and demosaicing (JDD). Although related to burst JDD, video JDD deserves its own treatment. In this work we present an empirical exploration of different design aspects of video joint denoising and demosaicing using neural networks. We compare recurrent and non-recurrent approaches and ex plore aspects such as type of propagated information in recurrent networks, moti on compensation, video stabilization, and network architecture. We found that re current networks with motion compensation achieve best results. Our work should serve as a strong baseline for future research in video JDD.

**********

Gallery Filter Network for Person Search

Lucas Jaffe, Avideh Zakhor; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 1684-1693

In person search, we aim to localize a query person from one scene in other gall ery scenes. The cost of this search operation is dependent on the number of gall ery scenes, making it beneficial to reduce the pool of likely scenes. We describ e and demonstrate the Gallery Filter Network (GFN), a novel module which can eff iciently discard gallery scenes from the search process, and benefit scoring for persons detected in remaining scenes. We show that the GFN is robust under a ra

nge of different conditions by testing on different retrieval sets, including cr
oss-camera, occluded, and low-resolution scenarios. In addition, we develop the
base SeqNeXt person search model, which improves and simplifies the original Seq
Net model. We show that the SeqNeXt+GFN combination yields significant performan
ce gains over other state-of-the-art methods on the standard PRW and CUHK-SYSU p
erson search datasets. To aid experimentation for this and other models, we prov
ide standardized tooling for the data processing and evaluation pipeline typical
ly used for person search research.
**********

A Continual Deepfake Detection Benchmark: Dataset, Methods, and Essentials
Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaope
ng Hong, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applicat
ions of Computer Vision (WACV), 2023, pp. 1339-1349
There have been emerging a number of benchmarks and techniques for the detection
 of deepfakes. However, very few works study the detection of incrementally appe
aring deepfakes in the real-world scenarios. To simulate the wild scenes, this p
aper suggests a continual deepfake detection benchmark (CDDB) over a new collect
ion of deepfakes from both known and unknown generative models. The suggested CD
DB designs multiple evaluations on the detection over easy, hard, and long seque
nce of deepfake tasks, with a set of appropriate measures. In addition, we explo
it multiple approaches to adapt multiclass incremental learning methods, commonl
y used in the continual visual recognition, to the continual deepfake detection
problem. We evaluate existing methods, including their adapted ones, on the prop
osed CDDB. Within the proposed benchmark, we explore some commonly known essenti
als of standard continual learning. Our study provides new insights on these ess
entials in the context of continual deepfake detection. The suggested CDDB is cl
early more challenging than the existing benchmarks, which thus offers a suitabl
e evaluation avenue to the future research. Both data and code are available at
https://github.com/Coral79/CDDB.
**********

Semantic Segmentation of Degraded Images Using Layer-Wise Feature Adjustor
Kazuki Endo, Masayuki Tanaka, Masatoshi Okutomi; Proceedings of the IEEE/CVF Win
ter Conference on Applications of Computer Vision (WACV), 2023, pp. 3205-3213
Semantic segmentation of degraded images is important for practical applications
 such as autonomous driving and surveillance systems. The degradation level, whi
ch represents the strength of degradation, is usually unknown in practice. There
fore, the semantic segmentation algorithm needs to take account of various level
s of degradation. In this paper, we propose a convolutional neural network of se
mantic segmentation which can cope with various levels of degradation. The propo
sed network is based on the knowledge distillation from a source network trained
 with only clean images. More concretely, the proposed network is trained to acq
uire multi-layer features keeping consistency with the source network, while adj
usting for various levels of degradation. The effectiveness of the proposed meth
od is confirmed for different types of degradations: JPEG distortion, Gaussian b
lur and salt&pepper noise. The experimental comparisons validate that the propos
ed network outperforms existing networks for semantic segmentation of degraded i
mages with various degradation levels.
**********

Addressing Feature Suppression in Unsupervised Visual Representations
Tianhong Li, Lijie Fan, Yuan Yuan, Hao He, Yonglong Tian, Rogerio Feris, Piotr I
ndyk, Dina Katabi; Proceedings of the IEEE/CVF Winter Conference on Applications
 of Computer Vision (WACV), 2023, pp. 1411-1420
Contrastive learning is one of the fastest growing research areas in machine lea
rning due to its ability to learn useful representations without labeled data. H
owever, contrastive learning is susceptible to feature suppression - i.e., it ma
y discard important information relevant to the task of interest, and learn irre
levant features. Past work has addressed this limitation via handcrafted data au
gmentations that eliminate irrelevant information. This approach however does no
t work across all datasets and tasks. Further, data augmentations fail in addres
sing feature suppression in multi-attribute classification when one attribute ca

n suppress features relevant to other attributes. In this paper, we analyze the objective function of contrastive learning and formally prove that it is vulnerable to feature suppression. We then present Predictive Contrastive Learning (PrCL), a framework for learning unsupervised representations that are robust to feature suppression. The key idea is to force the learned representation to predict the input, and hence prevent it from discarding important information. Extensive experiments verify that PrCL is robust to feature suppression and outperforms state-of-the-art contrastive learning methods on a variety of datasets and tasks.

**********

Guiding Users to Where To Give Color Hints for Efficient Interactive Sketch Colorization via Unsupervised Region Prioritization

Youngin Cho, Junsoo Lee, Soyoung Yang, Juntae Kim, Yeojeong Park, Haneol Lee, Mohammad Azam Khan, Daesik Kim, Jaegul Choo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1818-1827

Existing deep interactive colorization models have focused on ways to utilize various types of interactions, such as point-wise color hints, scribbles, or natural-language texts, as methods to reflect a user's intent at runtime. However, another approach, which actively informs the user of the most effective regions to give hints for sketch image colorization, has been under-explored. This paper proposes a novel model-guided deep interactive colorization framework that reduces the required amount of user interactions, by prioritizing the regions in a colorization model. Our method, called GuidingPainter, prioritizes these regions where the model most needs a color hint, rather than just relying on the user's manual decision on where to give a color hint. In our extensive experiments, we show that our approach outperforms existing interactive colorization methods in terms of the conventional metrics, such as PSNR and FID, and reduces required amount of interactions.

**********

Learning Incoherent Light Emission Steering From Metasurfaces Using Generative Models

Prasad P. Iyer, Saaketh Desai, Sadhvikas Addamane, Remi Dingreville, Igal Brener; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3770-3777

Spatiotemporal control over incoherent light sources is critically important for applications such as displays, remote sensing, clean energy, and illumination. Incoherent light emission made up of randomized wavefronts is incompatible with known beam steering techniques that rely on coherent electromagnetic wave interference. The emerging field of tunable dielectric metasurfaces consisting of sub-wavelength arrays of optical nanoresonators has recently enabled active re-direction of incoherent light (photoluminescence, PL) emission. This was achieved by illuminating (pumping) the metasurface with a pump laser reflecting off a programmable spatial light modulator (SLM) with sawtooth grating patterns as input. Achieving efficient beam steering requires the generation of optimal pump patterns programmed into the SLM to maximize the PL emitted towards a given direction. Given the innumerable possibilities and the lack of a theoretical physical framework to guide the exploration of pump patterns, we use an active learning algorithm running a closed loop optical experiment with a generative model to explore and optimize novel pump patterns. We achieve up to an order of magnitude enhancement in the steering efficiency by using pump patterns that are generated by a variational auto-encoder, with minimal number of experiments. The results presented in this paper highlight the unique ability of generative models and active learning to dramatically improve steering efficiency by finding novel optical pump patterns that are beyond human intuition. Our combination of advanced machine learning techniques driving closed loop nanophotonic experiments might pave the way to derive the underlying physics of emergent light-matter phenomena.

**********

Multimodal Vision Transformers With Forced Attention for Behavior Analysis

Tanay Agrawal, Michal Balazia, Philipp Müller, François Brémond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023,

pp. 3392-3402
Human behavior understanding requires looking at minute details in the large con text of a scene containing multiple input modalities. It is necessary as it allo ws the design of more human-like machines. While transformer approaches have sho wn great improvements, they face multiple challenges such as lack of data or bac kground noise. To tackle these, we introduce the Forced Attention (FAt) Transfor mer which utilize forced attention with a modified backbone for input encoding a nd a use of additional inputs. In addition to improving the performance on diffe rent tasks and inputs, the modification requires less time and memory resources. We provide a model for a generalised feature extraction for tasks concerning so cial signals and behavior analysis. Our focus is on understanding behavior in vi deos where people are interacting with each other or talking into the camera whi ch simulates the first person point of view in social interaction. FAt Transform ers are applied to two downstream tasks: personality recognition and body langua ge recognition. We achieve state-of-the-art results for Udiva v0.5, First Impres sions v2 and MPII Group Interaction datasets. We further provide an extensive ab lation study of the proposed architecture.
**********

Anomaly Detection in 3D Point Clouds Using Deep Geometric Descriptors
Paul Bergmann, David Sattlegger; Proceedings of the IEEE/CVF Winter Conference o n Applications of Computer Vision (WACV), 2023, pp. 2613-2623
We present a new method for the unsupervised detection of geometric anomalies in high-resolution 3D point clouds. In particular, we propose an adaptation of the established student-teacher anomaly detection framework to three dimensions. A student network is trained to match the output of a pretrained teacher network o n anomaly-free point clouds. When applied to test data, regression errors betwee n the teacher and the student allow reliable localization of anomalous structure s. To construct an expressive teacher network that extracts dense local geometri c descriptors, we introduce a novel self-supervised pretraining strategy. The te acher is trained by reconstructing local receptive fields and does not require a nnotations. Extensive experiments on the comprehensive MVTec 3D Anomaly Detectio n dataset highlight the effectiveness of our approach, which outperforms the exi sting methods by a large margin. Ablation studies show that our approach meets t he requirements of practical applications regarding performance, runtime, and me mory consumption.
**********

Self-Supervised Clustering Based on Manifold Learning and Graph Convolutional Ne tworks
Leonardo Tadeu Lopes, Daniel Carlos Guimarães Pedronette; Proceedings of the IEE E/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 563 4-5643
In spite of the huge advances in supervised learning, the common requirement for extensive labeled datasets represents a severe bottleneck. In this scenario, ot her learning paradigms capable of addressing the challenge associated with the s carcity of labeled data represent a relevant alternative solution. This paper pr esents a novel clustering method called Self-Supervised Graph Convolutional Clus tering (SGCC), which aims to exploit the strengths of different learning paradig ms, combining unsupervised, semi-supervised, and self-supervised perspectives. A n unsupervised manifold learning algorithm based on hypergraphs and ranking info rmation is used to provide more effective and global similarity information. The hypergraph structures allow identifying representative items for each cluster, which are used to derive a set of small but high confident clusters. Such cluste rs are taken as soft-labels for training a Graph Convolutional Network (GCN) in a semi-supervised classification task. Once trained in a self-supervised setting , the GCN is used to predict the cluster of remaining items. The proposed SGCC m ethod was evaluated both in image and citation networks datasets and compared wi th classic and recent clustering methods, obtaining high-effective results in al l scenarios.
**********

Orthogonal Transforms for Learning Invariant Representations in Equivariant Neur

al Networks
Jaspreet Singh, Chandan Singh, Ankur Rana; Proceedings of the IEEE/CVF Winter Co nference on Applications of Computer Vision (WACV), 2023, pp. 1523-1530
The convolutional layers of the standard convolutional neural networks (CNNs) ar e equivariant to translation. Recently, a new class of CNNs is introduced which is equivariant to other affine geometric transformations such as rotation and re flection by replacing the standard convolutional layer with the group convolutio nal layer or using the steerable filters in the convloutional layer. We propose to embed the 2D positional encoding which is invariant to rotation, reflection a nd translation using orthogonal polar harmonic transforms (PHTs) before flatteni ng the feature maps for fully-connected or classification layer in the equivaria nt CNN architecture. We select the PHTs among several invariant transforms, as t hey are very efficient in performance and speed. The proposed 2D positional enco ding scheme between the convolutional and fully-connected layers of the equivari ant networks is shown to provide significant improvement in performance on the r otated MNIST, CIFAR-10 and CIFAR-100 datasets.
**********

ViewCLR: Learning Self-Supervised Video Representation for Unseen Viewpoints
Srijan Das, Michael S. Ryoo; Proceedings of the IEEE/CVF Winter Conference on Ap plications of Computer Vision (WACV), 2023, pp. 5573-5583
Learning self-supervised video representation predominantly focuses on discrimin ating instances generated from simple data augmentation schemes. However, the le arned representation often fails to generalize over unseen camera viewpoints. To this end, we propose ViewCLR, that learns self-supervised video representation invariant to camera viewpoint changes. We introduce a viewpoint-generator that c an be considered as a learnable augmentation for any self-supervised pre-text ta sks, to generate latent viewpoint representation of a video. ViewCLR maximizes t he similarities between the representation of the latent viewpoint and that of t he original viewpoint, enabling the learned video encoder to generalize over uns een camera viewpoints. Experiments on cross-view benchmark datasets including NT U RGB+D dataset show that ViewCLR stands as a state-of-the-art viewpoint invaria nt self-supervised method.
**********

MRI Imputation Based on Fused Index- and Intensity-Registration
Jiyoon Shin, Jungwoo Lee; Proceedings of the IEEE/CVF Winter Conference on Appli cations of Computer Vision (WACV), 2023, pp. 1949-1958
3D MRI imaging is based on a number of imaging sequences such as T1, T2, T1ce, a nd Flair, and each of them is performed by a group of two-dimensional scans. In practical MRI, some scans are often missing while many medical applications requ ire a full set of scans. An MRI imputation method is presented, which synthesize s such missing scans. Key components in this method are the index registration a nd the intensity registration. The index registration models anatomical differen ces between two different scans in the same imaging sequence, and the intensity registration reflects the image contrast differences between two different scans of the same index. Two registration fields are learned to be invariant, and acc ordingly, allow two estimates of a missing scan, one within corresponding imagin g sequence and another along scan index; the two estimates are combined to yield the final synthesized scan. Experimental results highlight that the proposed me thod improves prevalent limitations existing in previous synthesis methods, blen ding both structural and contrast aspects and capturing subtle parts of the brai n. Quantitative results also show the superiority in various data sets, transiti ons, and measures.
**********

Masked Image Modeling Advances 3D Medical Image Analysis
Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Bala n, Kevin Brown; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1970-1980
Recently, masked image modeling (MIM) has gained considerable attention due to i ts capacity to learn from vast amounts of unlabeled data and has been demonstrat ed to be effective on a wide variety of vision tasks involving natural images. M

eanwhile, the potential of self-supervised learning in modeling 3D medical image s is anticipated to be immense due to the high quantities of unlabeled images, a nd the expense and difficulty of quality labels. However, MIM's applicability to medical images remains uncertain. In this paper, we demonstrate that masked ima ge modeling approaches can also advance 3D medical images analysis in addition t o natural images. We study how masked image modeling strategies leverage perform ance from the viewpoints of 3D medical image segmentation as a representative do wnstream task: i) when compared to naive contrastive learning, masked image mode ling approaches accelerate the convergence of supervised training even faster (1 .40x) and ultimately produce a higher dice score; ii) predicting raw voxel value s with a high masking ratio and a relatively smaller patch size is non-trivial s elf-supervised pretext-task for medical images modeling; iii) a lightweight deco der or projection head design for reconstruction is powerful for masked image mo deling on 3D medical images which speeds up training and reduce cost; iv) finall y, we also investigate the effectiveness of MIM methods under different practica l scenarios where different image resolutions and labeled data ratios are applie d. Anonymized codes are available at https://anonymous.4open.science/r/MIM-Med3D .

**********

PreViTS: Contrastive Pretraining With Video Tracking Supervision

Brian Chen, Ramprasaath R. Selvaraju, Shih-Fu Chang, Juan Carlos Niebles, Nikhil Naik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1560-1570

Videos are a rich source for self-supervised learning (SSL) of visual representa tions due to the presence of natural temporal transformations of objects. Howeve r, current methods typically randomly sample video clips for learning, which res ults in an imperfect supervisory signal. In this work, we propose PreViTS, an SS L framework that utilizes an unsupervised tracking signal for selecting clips co ntaining the same object, which helps better utilize temporal transformations of objects. PreViTS further uses the tracking signal to spatially constrain the fr ame regions to learn from and trains the model to locate meaningful objects by p roviding supervision on Grad-CAM attention maps. To evaluate our approach, we tr ain a momentum contrastive (MoCo) encoder on VGG-Sound and Kinetics-400 datasets with PreViTS. Training with PreViTS outperforms representations learnt by contr astive strategy alone on video downstream tasks, obtaining state-of-the-art perf ormance on action classification. PreViTS helps learn feature representations th at are more robust to changes in background and context, as seen by experiments on datasets with background changes. Learning from large-scale videos with PreVi TS could lead to more accurate and robust visual feature representations.

**********

Foreground Guidance and Multi-Layer Feature Fusion for Unsupervised Object Disco very With Transformers

Zhiwei Lin, Zengyu Yang, Yongtao Wang; Proceedings of the IEEE/CVF Winter Confer ence on Applications of Computer Vision (WACV), 2023, pp. 4043-4053

Unsupervised object discovery (UOD) has recently shown encouraging progress with the adoption of pre-trained Transformer features. However, current methods base d on Transformers mainly focus on designing the localization head (e.g., seed se lection-expansion and normalized cut) and overlook the importance of improving T ransformer features. In this work, we handle UOD task from the perspective of fe ature enhancement and propose FOReground guidance and MUlti-LAyer feature fusion for unsupervised object discovery, dubbed FORMULA. Firstly, we present a foregr ound guidance strategy with an off-the-shelf UOD detector to highlight the foreg round regions on the feature maps and then refine object locations in an iterati ve fashion. Moreover, to solve the scale variation issues in object detection, w e design a multi-layer feature fusion module that aggregates features responding to objects at different scales. The experiments on VOC07, VOC12, and COCO_20k s how that the proposed FORMULA achieves new state-of-the-art results on unsupervi sed object discovery. The code will be released at https://github.com/VDIGPKU/FO RMULA.

**********

Gait Recognition Using 3-D Human Body Shape Inference
Haidong Zhu, Zhaoheng Zheng, Ram Nevatia; Proceedings of the IEEE/CVF Winter Con
ference on Applications of Computer Vision (WACV), 2023, pp. 909-918
Gait recognition, which identifies individuals based on their walking patterns,
is an important biometric technique since it can be observed from a distance and
 does not require the subject's cooperation. Recognizing a person's gait is diff
icult because of the appearance variants in human silhouette sequences produced
by varying viewing angles, carrying objects, and clothing. Recent research has p
roduced a number of ways for coping with these variants. In this paper, we prese
nt the usage of inferring 3-D body shapes distilled from limited images, which a
re, in principle, invariant to the specified variants. Inference of 3-D shape is
 a difficult task, especially when only silhouettes are provided in a dataset. W
e provide a method for learning 3-D body inference from silhouettes by transferr
ing knowledge from 3-D shape prior from RGB photos. We use our method on multipl
e existing state-of-the-art gait baselines and obtain consistent improvements fo
r gait identification on two public datasets, CASIA-B and OUMVLP, on several var
iants and settings, including a new setting of novel views not seen during train
ing.
**********
Delving Into Masked Autoencoders for Multi-Label Thorax Disease Classification
Junfei Xiao, Yutong Bai, Alan Yuille, Zongwei Zhou; Proceedings of the IEEE/CVF
Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3588-3600
Vision Transformer (ViT) has become one of the most popular neural architectures
 due to its simplicity, scalability, and compelling performance in multiple visi
on tasks. However, since the scales of medical datasets are relatively small, Vi
T has shown inferior performance on medical datasets even after pre-trained on I
mageNet. In this paper, we unleash the potential of ViT by pre-training on 266,3
40 unlabeled chest X-rays. Specifically, we explore Masked Autoencoders (MAE) wh
ose task is to reconstruct missing pixels from a small proportion of each image
and figure out a strong recipe for pre-training MAE and fine-tuning on chest X-r
ay datasets, revealing that medical reconstruction needs a much smaller proporti
on of an image than natural images (10% vs. 25%) and a more moderate RandomResiz
edCrop cropping range than natural images (0.5 1.0 vs. 0.2 1.0). With our recipe
, ViT-S shows competitive results with the state-of-the-art CNN model (DenseNet-
121) on three public chest X-ray datasets and 2.5x faster pre-training on the NI
H ChestX-ray14 dataset and CheXpert. To the best of our knowledge, we are the fi
rst to make vanilla ViT achieve state-of-the-art performance on chest X-ray data
sets. We hope that this study can direct future research on the application of T
ransformers to a larger variety of medical imaging tasks. Code will be made avai
lable.
**********
Dissecting Deep Metric Learning Losses for Image-Text Retrieval
Hong Xuan, Xi (Stephen) Chen; Proceedings of the IEEE/CVF Winter Conference on A
pplications of Computer Vision (WACV), 2023, pp. 2164-2173
Visual-Semantic Embedding (VSE) is a prevalent approach in image-text retrieval
by learning a joint embedding space between the image and language modalities wh
ere semantic similarities would be preserved. The triplet loss with hard-negativ
e mining has become the de-facto objective for most VSE methods. Inspired by rec
ent progress in deep metric learning (DML) in the image domain which gives rise
to new loss functions that outperform triplet loss, in this paper we revisit the
 problem of finding better objectives for VSE in image-text matching. Despite so
me attempts in designing losses based on gradient movement, most DML losses are
defined empirically in the embedding space. Instead of directly applying these l
oss functions which may lead to sub-optimal gradient updates in model parameters
, in this paper we present a novel Gradient-based Objective AnaLysis framework,
or GOAL, to systematically analyze the combinations and reweighting of the gradi
ents in existing DML functions. With the help of this analysis framework, we fur
ther propose a new family of objectives in the gradient space exploring differen
t gradient combinations. In the event that the gradients are not integrable to a
 valid loss function, we implement our proposed objectives such that they would

directly operate in the gradient space instead of on the losses in the embedding space. Comprehensive experiments have demonstrated that our novel objectives have consistently improved performance over baselines across different visual/text features and model frameworks. We also showed the generalizability of the GOAL framework by extending it to other models using triplet family losses including vision-language model with heavy cross-modal interactions and have achieved state-of-the-art results on the image-text retrieval tasks on COCO and Flick30K.
**********

BoxMask: Revisiting Bounding Box Supervision for Video Object Detection

Khurram Azeem Hashmi, Alain Pagani, Didier Stricker, Muhammad Zeshan Afzal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2030-2040

We present a new, simple yet effective approach to uplift video object detection. We observe that prior works operate on instance-level feature aggregation that imminently neglects the refined pixel-level representation, resulting in confusion among objects sharing similar appearance or motion characteristics. To address this limitation, we pro- pose BoxMask, which effectively learns discriminative representations by incorporating class-aware pixel-level information. We simply consider bounding box-level annotations as a coarse mask for each object to supervise our method. The proposed module can be effortlessly integrated into any region-based detector to boost detection. Extensive experiments on ImageNet VID and EPIC KITCHENS datasets demonstrate consistent and significant improvement when we plug our BoxMask module into numerous recent state-of-the-art methods. The code will be available at https://github.com/khurramHashmi/BoxMask.
**********

A Simple and Powerful Global Optimization for Unsupervised Video Object Segmentation

Georgy Ponimatkin, Nermin Samet, Yang Xiao, Yuming Du, Renaud Marlet, Vincent Lepetit; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5892-5903

We propose a simple, yet powerful approach for unsupervised object segmentation in videos. We introduce an objective function whose minimum represents the mask of the main salient object over the input sequence. It only relies on independent image features and optical flows, which can be obtained using off-the-shelf self-supervised methods. It scales with the length of the sequence with no need for superpixels or sparsification, and it generalizes to different datasets without any specific training. This objective function can actually be derived from a form of spectral clustering applied to the entire video. Our method achieves on-par performance with the state of the art on standard benchmarks (DAVIS2016, SegTrack-v2, FBMS59), while being conceptually and practically much simpler.
**********

Randomness Is the Root of All Evil: More Reliable Evaluation of Deep Active Learning

Yilin Ji, Daniel Kaestner, Oliver Wirth, Christian Wressnegger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3943-3952

Using deep neural networks for active learning (AL) poses significant challenges for the stability and the reproducibility of experimental results. Inconsistent settings continue to be the root causes for contradictory conclusions and in worst cases, for incorrect appraisal of methods. Our community is in search of a unified framework for exhaustive and fair evaluation of deep active learning. In this paper, we provide just such a framework, one which is built upon systematically fixing, containing and interpreting sources of randomness. We isolate different influence factors, such as neural-network initialization or hardware specifics, to assess their impact on the learning performance. We then use our framework to analyze the effects of basic AL settings, such as the query-batch size and the use of subset selection, and different datasets on AL performance. Our findings enable us to derive specific recommendations for the reliable evaluation of deep active learning, thus helping advance the community toward a more normative evaluation of results.

**********
Visually Explaining 3D-CNN Predictions for Video Classification With an Adaptive Occlusion Sensitivity Analysis
Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, Kazuhiro Fukui; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1513-1522
This paper proposes a method for visually explaining the decision-making process of 3D convolutional neural networks (CNN) with a temporal extension of occlusion sensitivity analysis. The key idea here is to occlude a specific volume of data by a 3D mask in an input 3D temporal-spatial data space and then measure the change degree in the output score. The occluded volume data that produces a larger change degree is regarded as a more critical element for classification. However, while the occlusion sensitivity analysis is commonly used to analyze single image classification, it is not so straightforward to apply this idea to video classification as a simple fixed cuboid cannot deal with the motions. To this end, we adapt the shape of a 3D occlusion mask to complicated motions of target objects. Our flexible mask adaptation is performed by considering the temporal continuity and spatial co-occurrence of the optical flows extracted from the input video data. We further propose to approximate our method by using the first-order partial derivative of the score with respect to an input image to reduce its computational cost. We demonstrate the effectiveness of our method through various and extensive comparisons with the conventional methods in terms of the deletion/insertion metric and the pointing metric on the UCF-101. The code is available at: https://github.com/uchiyama33/AOSA.
**********
Self-Supervised Learning With Local Contrastive Loss for Detection and Semantic Segmentation
Ashraful Islam, Benjamin Lundell, Harpreet Sawhney, Sudipta N. Sinha, Peter Morales, Richard J. Radke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5624-5633
We present a self-supervised learning (SSL) method suitable for semi-global tasks such as object detection and semantic segmentation. We enforce local consistency between self-learned features, representing corresponding image locations of transformed versions of the same image, by minimizing a pixel-level local contrastive (LC) loss during training. LC-loss can be added to existing self-supervised learning methods with minimal overhead. We evaluate our SSL approach on two downstream tasks -- object detection and semantic segmentation, using COCO, PASCAL VOC, and CityScapes datasets. Our method outperforms the existing state-of-the-art SSL approaches by 1.9% on COCO object detection, 1.4% on PASCAL VOC detection, and 0.6% on CityScapes segmentation.
**********
Concept Correlation and Its Effects on Concept-Based Models
Lena Heidemann, Maureen Monnet, Karsten Roscher; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4780-4788
Concept-based learning approaches for image classification, such as Concept Bottleneck Models, aim to enable interpretation and increase robustness by directly learning high-level concepts which are used for predicting the main class. They achieve competitive test accuracies compared to standard end-to-end models. However, with multiple concepts per image and binary concept annotations (without concept localization), it is not evident if the output of the concept model is truly based on the predicted concepts or other features in the image. Additionally, high correlations between concepts would allow a model to predict a concept with high test accuracy by simply using a correlated concept as a proxy. In this paper, we analyze these correlations between concepts in the CUB and GTSRB datasets and propose methods beyond test accuracy for evaluating their effects on the performance of a concept-based model trained on this data. To this end, we also perform a more detailed analysis on the effects of concept correlation using synthetically generated datasets of 3D shapes. We see that high concept correlation increases the risk of a model's inability to distinguish these concepts. Yet simple techniques, like loss weighting, show promising initial results for mitigati

ng this issue.
**********
A Neural Video Codec With Spatial Rate-Distortion Control
Noor Fathima, Jens Petersen, Guillaume Sautière, Auke Wiggers, Reza Pourreza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5365-5374
Neural video compression algorithms are nearly competitive with hand-crafted codecs in terms of rate-distortion performance and subjective quality. However, many neural codecs are inflexible black boxes, and give users little to no control over the reconstruction quality and bitrate. In this work, we present a flexible neural video codec that combines ideas from variable-bitrate codecs and region-of-interest-based coding. By conditioning our model on a global rate-distortion tradeoff parameter and a region-of-interest (ROI) mask, we obtain fine control over the per-frame bitrate and the reconstruction quality in the ROI. The resulting codec enables practical use cases such as coding under bitrate constraints with fixed ROI quality, while taking a negligible hit in overall rate-distortion performance. We find that our codec is best utilized when the sequence contains complex motion, such as scenes with camera panning or sports videos, where we substantially outperform non-ROI codecs in the region of interest with BD-rate savings exceeding 60% in some cases.
**********
Planar Object Tracking via Weighted Optical Flow
Jonáš Šerých, Ji■í Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1593-1602
We propose WOFT - a novel method for planar object tracking that estimates a full 8 degrees-of-freedom pose, i.e., the homography w.r.t. a reference view. The method uses a novel module that leverages dense optical flow and assigns a weight to each optical flow correspondence, estimating a homography by weighted least squares in a fully differentiable manner. The trained module assigns zero weights to incorrect correspondences (outliers) in most cases, making the method robust and eliminating the need of the typically used non-differentiable robust estimators like RANSAC. The proposed weighted optical flow tracker (WOFT) achieves state-of-the-art performance on two benchmarks, POT-210 and POIC, tracking consistently well across a wide range of scenarios.
**********
Dance Style Transfer With Cross-Modal Transformer
Wenjie Yin, Hang Yin, Kim Baraka, Danica Kragic, Mårten Björkman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5058-5067
We present CycleDance, a dance style transfer system to transform an existing motion clip in one dance style to a motion clip in another dance style while attempting to preserve motion context of the dance. Our method extends an existing CycleGAN architecture for modeling audio sequences and integrates multimodal transformer encoders to account for music context. We adopt sequence length-based curriculum learning to stabilize training. Our approach captures rich and long-term intra-relations between motion frames, which is a common challenge in motion transfer and synthesis work. We further introduce new metrics for gauging transfer strength and content preservation in the context of dance movements. We perform an extensive ablation study as well as a human study including 30 participants with 5 or more years of dance experience. The results demonstrate that CycleDance generates realistic movements with the target style, significantly outperforming the baseline CycleGAN on naturalness, transfer strength, and content preservation.
**********
Interpreting Disparate Privacy-Utility Tradeoff in Adversarial Learning via Attribute Correlation
Likun Zhang, Yahong Chen, Ang Li, Binghui Wang, Yiran Chen, Fenghua Li, Jin Cao, Ben Niu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4701-4709
Adversarial learning is commonly used to extract latent data representations whi

ch are expressive to predict the target attribute but indistinguishable in the p
rivacy attribute. However, whether they can achieve an expected privacy-utility
tradeoff is of great uncertainty. In this paper, we posit it is the complex inte
raction between different attributes in the training set that causes disparate t
radeoff results. We first formulate the measurement of utility, privacy and thei
r tradeoff in adversarial learning. Then we propose the metrics of Statistical R
eliability (SR) and Feature Reliability (FR) to quantify the relationship betwee
n attributes. Specifically, SR reflects the co-occurrence sampling bias of the j
oint distribution between two attributes. Beyond the explicit dependence, FR exp
loits the intrinsic interaction one attribute exerts on the other via exploring
the representation disentanglement. We validate the metrics in an adversarial le
arning scheme on CelebA and LFW dataset with a suite of target-privacy attribute
 pairs. Experiments demonstrate the strong correlations between the metrics and
utility, privacy and their tradeoff. We further conclude how to use SR and FR as
 a guide to the selection of the privacy-utility tradeoff parameter.
**********

Control-NeRF: Editable Feature Volumes for Scene Rendering and Manipulation
Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, Gerard Pons-Moll
; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi
on (WACV), 2023, pp. 4340-4350
We present Control-NeRF, a method for performing flexible, 3D-aware image conten
t manipulation while enabling high-quality novel view synthesis, from a set of p
osed input images. NeRF-based approaches are effective for novel view synthesis,
 however such models memorize the radiance for every point in a scene within a n
eural network. Since these models are scene-specific and lack a 3D scene represe
ntation, classical editing such as shape manipulation, or combining scenes is no
t possible. While there are some recent hybrid approaches that combine NeRF with
 external scene representations such as sparse voxels, planes, hash tables, etc.
, they focus mostly on efficiency and don't explore the scene editing and manipu
lation capabilities of hybrid approaches. With the aim of exploring controllable
 scene representations for novel view synthesis, our model couples learnt scene-
specific 3D feature volumes with a general NeRF rendering network. We can genera
lize to novel scenes by optimizing only the scene-specific 3D feature volume, wh
ile keeping the parameters of the rendering network fixed. Since the feature vol
umes are independent of the rendering model, we can manipulate and combine scene
s by editing their corresponding feature volumes. The edited volume can then be
plugged into the rendering model to synthesise high-quality novel views. We demo
nstrate scene manipulations including: scene mixing; applying rigid and non-rigi
d transformations; inserting, moving and deleting objects in a scene; while prod
ucing photo-realistic novel-view synthesis results.
**********

Towards Equivariant Optical Flow Estimation With Deep Learning
Stefano Savian, Pietro Morerio, Alessio Del Bue, Andrea A. Janes, Tammam Tillo;
Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision
 (WACV), 2023, pp. 5088-5097
Methods for Optical Flow (OF) estimation based on Deep Learning have considerabl
y improved traditional approaches in challenging and realistic conditions. Howev
er, data-driven approaches can inherently be biased, leading to unexpected under
-performance in real application scenarios. In this paper, we first observe that
 the OF estimation accuracy varies with motion direction, and name this phenomen
on 'OF sign imbalance'. The sign imbalance cannot be assessed by means of the en
dpoint-error (EPE), the typical training and evaluation metric for Deep Optical
Flow estimators. This paper tackles this issue by proposing a new metric to asse
ss the sign imbalance, which is compared to the endpoint-error. We provide an ex
tensive evaluation of the sign imbalance for the state-of-the-art optical flow e
stimators. Based on the evaluation, we propose two strategies to mitigate the ph
enomenon, i) by constraining the model estimations during inference, and, ii) by
 constraining the loss function during training. Testing and training code is av
ailable at: www.github.com/stsavian/equivariant_of_estimation.
**********

Hyperblock Floating Point: Generalised Quantization Scheme for Gradient and Inference Computation
Marcelo Gennari do Nascimento, Victor Adrian Prisacariu, Roger Fawcett, Martin Langhammer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6364-6373
Prior quantization methods focus on producing networks for fast and lightweight inference. However, the cost of unquantised training is overlooked, despite requiring significantly more time and energy than inference. We present a method for quantizing convolutional neural networks for efficient training. Quantizing gradients is challenging because it requires higher granularity and their values span a wider range than the weight and feature maps. We propose an extension of the Channel-wise Block Floating Point format that allows for quick gradient computation, using a minimal amount of quantization time. This is achieved through sharing an exponent across both depth and batch dimensions in order to quantize tensors once and reuse them during backpropagation. We test our method using standard models such as AlexNet, VGG, and ResNet, on the CIFAR10, SVHN and ImageNet datasets. We show no loss of accuracy when quantizing AlexNet weights, activations and gradients to only 4 bits training ImageNet.
**********
FeTrIL: Feature Translation for Exemplar-Free Class-Incremental Learning
Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, Bertrand Delezoide; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3911-3920
Exemplar-free class-incremental learning is very challenging due to the negative effect of catastrophic forgetting. A balance between stability and plasticity of the incremental process is needed in order to obtain good accuracy for past as well as new classes. Existing exemplar-free class-incremental methods focus either on successive fine tuning of the model, thus favoring plasticity, or on using a feature extractor fixed after the initial incremental state, thus favoring stability. We introduce a method which combines a fixed feature extractor and a pseudo-features generator to improve the stability-plasticity balance. The generator uses a simple yet effective geometric translation of new class features to create representations of past classes, made of pseudo-features. The translation of features only requires the storage of the centroid representations of past classes to produce their pseudo-features. Actual features of new classes and pseudo-features of past classes are fed into a linear classifier which is trained incrementally to discriminate between all classes. The incremental process is much faster with the proposed method compared to mainstream ones which update the entire deep model. Experiments are performed with three challenging datasets, and different incremental settings. A comparison with ten existing methods shows that our method outperforms the others in most cases. FeTrIL code is available at https://github.com/GregoirePetit/FeTrIL
**********
An Embedding-Dynamic Approach to Self-Supervised Learning
Suhong Moon, Domas Buracas, Seunghyun Park, Jinkyu Kim, John Canny; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2750-2758
A number of recent self-supervised learning methods have shown impressive performance on image classification and other tasks. A somewhat bewildering variety of techniques have been used, not always with a clear understanding of the reasons for their benefits, especially when used in combination. Here we treat the embeddings of images as point particles and consider model optimization as a dynamic process on this system of particles. Our dynamic model combines an attractive force for similar images, a locally dispersive force to avoid local collapse, and a global dispersive force to achieve a globally-homogeneous distribution of particles. The dynamic perspective highlights the advantage of using a delayed-parameter image embedding (a la BYOL) together with multiple views of the same image. It also uses a purely-dynamic local dispersive force (Brownian motion) that shows improved performance over other methods and does not require knowledge of other particle coordinates. The method is called MSBReg which stands for (i) a Mul

tiview centroid loss, which applies an attractive force to pull different image view embeddings toward their centroid, (ii) a Singular value loss, which pushes the particle system toward spatially homogeneous density, (iii) a Brownian diffusive loss. We evaluate downstream classification performance of MSBReg on ImageNet as well as transfer learning tasks including fine-grained classification, multi-class object classification, object detection, and instance segmentation. In addition, we also show that applying our regularization term to other methods further improves their performance and stabilize the training by preventing a mode collapse.

**********

Universal Deep Image Compression via Content-Adaptive Optimization With Adapters
Koki Tsubota, Hiroaki Akutsu, Kiyoharu Aizawa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2529-2538
Deep image compression performs better than conventional codecs, such as JPEG, on natural images. However, deep image compression is learning-based and encounters a problem: the compression performance deteriorates significantly for out-of-domain images. In this study, we highlight this problem and address a novel task: universal deep image compression. This task aims to compress images belonging to arbitrary domains, such as natural images, line drawings, and comics. To address this problem, we propose a content-adaptive optimization framework; this framework uses a pre-trained compression model and adapts the model to a target image during compression. Adapters are inserted into the decoder of the model. For each input image, our framework optimizes the latent representation extracted by the encoder and the adapter parameters in terms of rate-distortion. The adapter parameters are additionally transmitted per image. For the experiments, a benchmark dataset containing uncompressed images of four domains (natural images, line drawings, comics, and vector arts) is constructed and the proposed universal deep compression is evaluated. Finally, the proposed model is compared with non-adaptive and existing adaptive compression models. The comparison reveals that the proposed model outperforms these. The code and dataset are publicly available at https://github.com/kktsubota/universal-dic.

**********

Semantic Guided Latent Parts Embedding for Few-Shot Learning
Fengyuan Yang, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5447-5457
The ability of few-shot learning (FSL) is a basic requirement of intelligent agent learning in the open visual world. However, existing deep learning systems rely too heavily on large numbers of training samples, making it hard to learn new categories efficiently from limited size of training data. Two key challenges of FSL are insufficient comprehension and imperfect modeling of the few-shot novel class. For insufficient visual comprehension, semantic knowledge which is information from other modalities can help replenish the understanding of novel classes. But even so, most works still suffer from the second challenge because the single global class prototype they adopted is extremely unstable and imperfect given the larger intra-class variation and harder inter-class discrimination in FSL scenario. Thus, we propose to represent each class by its several different parts with the help of class semantic knowledge. Since we can never pre-define parts for unknown novel classes, we embed them in a latent manner. Concretely, we train a generator that takes the class semantic knowledge as input and outputs several filters of class-specific semantic latent parts. By applying each part filter, our model can pay attention to corresponding local regions containing each part. At the inference stage, the classification is conducted by comparing the similarities between those parts. Experiments on several FSL benchmarks demonstrate the effectiveness of our proposed method and show its potential to go beyond class recognition to class understanding. Furthermore, we also find when semantic knowledge is more visualized and customized, it will be more helpful in the FSL task.

**********

QMagFace: Simple and Accurate Quality-Aware Face Recognition
Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner

, Kiran Raja, Arjan Kuijper; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3484-3494
In this work, we propose QMagFace, a simple and effective face recognition solution (QMagFace) that combines a quality-aware comparison score with a recognition model based on a magnitude-aware angular margin loss.The proposed approach includes model-specific face image qualities in the comparison process to enhance the recognition performance under unconstrained circumstances. Exploiting the linearity between the qualities and their comparison scores induced by the utilized loss, our quality-aware comparison function is simple and highly generalizable. The experiments conducted on several face recognition databases and benchmarks demonstrate that the introduced quality-awareness leads to consistent improvements in the recognition performance. Moreover, the proposed QMagFace approach performs especially well under challenging circumstances, such as cross-pose, cross-age, or cross-quality. Consequently, it leads to state-of-the-art performances on several face recognition benchmarks, such as 98.50% on AgeDB, 83.95% on XQLFQ, and 98.74% on CFP-FP. The code for QMagFace is publicly available.
**********

Learning Graph Variational Autoencoders With Constraints and Structured Priors for Conditional Indoor 3D Scene Generation
Aditya Chattopadhyay, Xi Zhang, David Paul Wipf, Himanshu Arora, René Vidal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 785-794
We present a graph variational autoencoder with a structured prior for generating the layout of indoor 3D scenes. Given the room type (e.g., living room or library) and the room layout (e.g., room elements such as floor and walls), our architecture generates a collection of objects (e.g., furniture items such as sofa, table and chairs) that is consistent with the room type and layout. This is a challenging problem because the generated scene needs to satisfy multiple constrains, e.g., each object should lie inside the room and two objects should not occupy the same volume. To address these challenges, we propose a deep generative model that encodes these relationships as soft constraints on an attributed graph (e.g., the nodes capture attributes of room and furniture elements, such as shape, class, pose and size, and the edges capture geometric relationships such as relative orientation). The architecture consists of a graph encoder that maps the input graph to a structured latent space, and a graph decoder that generates a furniture graph, given a latent code and the room graph. The latent space is modeled with autoregressive priors, which facilitates the generation of highly structured scenes. We also propose an efficient training procedure that combines matching and constrained learning. Experiments on the 3D-FRONT dataset show that our method produces scenes that are diverse and are adapted to the room layout.
**********

DeepPrivacy2: Towards Realistic Full-Body Anonymization
Håkon Hukkelås, Frank Lindseth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1329-1338
Generative Adversarial Networks (GANs) are widely adapted for anonymization of human figures. However, current state-of-the-art limit anonymization to the task of face anonymization. In this paper, we propose a novel anonymization framework (DeepPrivacy2) for realistic anonymization of human figures and faces. We introduce a new large and diverse dataset for human figure synthesis, which significantly improves image quality and diversity of generated images. Furthermore, we propose a style-based GAN that produces high quality, diverse and editable anonymizations. We demonstrate that our full-body anonymization framework provides stronger privacy guarantees than previously proposed methods.
**********

Back to MLP: A Simple Baseline for Human Motion Prediction
Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4809-4819
This paper tackles the problem of human motion prediction, consisting in forecasting future body poses from historically observed sequences. State-of-the-art ap

proaches provide good results, however, they rely on deep learning architectures of arbitrary complexity, such as Recurrent Neural Networks(RNN), Transformers or Graph Convolutional Networks(GCN), typically requiring multiple training stages and more than 2 million parameters. In this paper, we show that, after combining with a series of standard practices, such as applying Discrete Cosine Transform (DCT), predicting residual displacement of joints and optimizing velocity as an auxiliary loss, a light-weight network based on multi-layer perceptrons (MLPs) with only 0.14 million parameters can surpass the state-of-the-art performance. An exhaustive evaluation on the Human3.6M, AMASS, and 3DPW datasets shows that our method, named siMLPe, consistently outperforms all other approaches. We hope that our simple method could serve as a strong baseline for the community and allow re-thinking of the human motion prediction problem. The code is publicly available at https://github.com/dulucas/siMLPe.

**********

Content-Based Music-Image Retrieval Using Self- and Cross-Modal Feature Embedding Memory

Takayuki Nakatsuka, Masahiro Hamasaki, Masataka Goto; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2174-2184

This paper describes a method based on deep metric learning for content-based cross-modal retrieval of a piece of music and its representative image (i.e., a music audio signal and its cover art image). We train music and image encoders so that the embeddings of a positive music-image pair lie close to each other, while those of a random pair lie far from each other, in a shared embedding space. Furthermore, we propose a mechanism called self- and cross-modal feature embedding memory, which stores both the music and image embeddings of any previous iterations in memory and enables the encoders to mine informative pairs for training. To perform such training, we constructed a dataset containing 78,325 music-image pairs. We demonstrate the effectiveness of the proposed mechanism on this dataset: specifically, our mechanism outperforms baseline methods by 1.93 3.38 times for the mean reciprocal rank, 2.19 3.56 times for recall@50, and 528 891 ranks for the median rank.

**********

Creating a Forensic Database of Shoeprints From Online Shoe-Tread Photos

Samia Shafique, Bailey Kong, Shu Kong, Charless Fowlkes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 858-868

Shoe tread impressions are one of the most common types of evidence left at crime scenes. However, the utility of such evidence is limited by the lack of databases of footwear prints that cover the large and growing number of distinct shoe models. Moreover, the database is preferred to contain the 3D shape, or depth, of shoe-tread photos so as to allow for extracting shoeprints to match a query (crime-scene) print. We propose to address this gap by leveraging shoe-tread photos collected by online retailers. The core challenge is to predict depth maps for these photos. As they do not have ground-truth 3D shapes allowing for training depth predictors, we exploit synthetic data that does. We develop a method termed ShoeRinsics that learns to predict depth by leveraging a mix of fully supervised synthetic data and unsupervised retail image data. In particular, we find domain adaptation and intrinsic image decomposition techniques effectively mitigate the synthetic-real domain gap and yield significantly better depth prediction. To validate our method, we introduce 2 validation sets consisting of shoe-tread image and print pairs and define a benchmarking protocol to quantify the quality of predicted depth. On this benchmark, ShoeRinsics outperforms existing methods of depth prediction and synthetic-to-real domain adaptation.

**********

SCTS: Instance Segmentation of Single Cells Using a Transformer-Based Semantic-Aware Model and Space-Filling Augmentation

Yating Zhou, Wenjing Li, Ge Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5944-5953

Instance segmentation of single cells from microscopy images is critical to quan

titative analysis of their spatial and morphological features for many important biomedical applications, such as disease diagnosis and drug screening. However, the high densities, tight contacts, and weak boundaries of the cells pose substantial technical challenges. To overcome these challenges, we have developed a new instance segmentation model, which we refer to as single-cell Transformer segmenter (SCTS). It utilizes a Swin Transformer as its backbone, combining the global modeling capabilities of a Transformer and the local modeling capabilities of a convolutional neural network (CNN) to ensure model adaptability to different cell sizes, shapes, and textures. It also embeds a three-class (background, cell interior, and cell boundary) semantic segmentation branch to classify pixels and to provide semantic features for downstream tasks. The prediction of boundary semantics improves boundary awareness, and the differentiation between foreground and background semantics improves segmentation integrity in regions with weak signals. To reduce the need for annotated training data, we have developed an augmentation strategy that randomly fills instances of single cells into open spaces of training images. Experiments show that our model outperforms several state-of-the-art models on the LIVECell dataset and an in-house dataset. The code and dataset of this work are openly accessible at https://github.com/cbmi-group/SCTS.

**********

SIRA: Relightable Avatars From a Single Image

Pol Caselles, Eduard Ramon, Jaime Garcia, Xavier Giro-i-Nieto, Francesc Moreno-Noguer, Gil Triginer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 775-784

Recovering the geometry of a human head from a single image, while factorizing the materials and illumination is a severely ill-posed problem that requires prior information to be solved. Methods based on 3D Morphable Models (3DMM), and their combination with differentiable renderers, have shown promising results. However, the expressiveness of 3DMMs is limited, and they typically yield over-smoothed and identity-agnostic 3D shapes limited to the face region. Highly accurate full head reconstructions have recently been obtained with neural fields that parameterize the geometry using multilayer perceptrons. The versatility of these representations has also proved effective for disentangling geometry, materials and lighting. However, these methods require several tens of input images. In this paper, we introduce SIRA, a method which, from a single image, reconstructs human head avatars with high fidelity geometry and factorized lights and surface materials. Our key ingredients are two data-driven statistical models based on neural fields that resolve the ambiguities of single-view 3D surface reconstruction and appearance factorization. Experiments show that SIRA obtains state of the art results in 3D head reconstruction while at the same time it successfully disentangles the global illumination, and the diffuse and specular albedos. Furthermore, our reconstructions are amenable to physically-based appearance editing and head model relighting.

**********

Performance Comparison of DVS Data Spatial Downscaling Methods Using Spiking Neural Networks

Amélie Gruel, Jean Martinet, Bernabé Linares-Barranco, Teresa Serrano-Gotarredona; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6494-6502

Dynamic Vision Sensors (DVS) are an unconventional type of camera that produces sparse and asynchronous event data, which has recently led to a strong increase in its use for computer vision tasks namely in robotics. Embedded systems face limitations in terms of energy resources, memory, computational power, and communication bandwidth. Hence, this application calls for a way to reduce the amount of data to be processed while keeping the relevant information for the task at hand. We thus believe that a formal definition of event data reduction methods will provide a step further towards sparse data processing. The contributions of this paper are twofold: we introduce two complementary neuromorphic methods based on Spiking Neural Networks for DVS data spatial reduction, which is to best of our knowledge the first proposal of neuromorphic event data reduction; then we s

tudy for each method the trade-off between the amount of information kept after reduction, the performance of gesture classification after reduction and their capacity to handle events in real time. We demonstrate here that the proposed SNN-based methods outperform existing methods in a classification task for most dividing factors and are significantly better at handling data in real time, and make therefore the optimal choice for fully-integrated energy-efficient event data reduction running dynamically on a neuromorphic platform. Our code is publicly available online at: https://github.com/amygruel/EvVisu.

**********

EmbryosFormer: Deformable Transformer and Collaborative Encoding-Decoding for Embryos Stage Development Classification

Tien-Phat Nguyen, Trong-Thang Pham, Tri Nguyen, Hieu Le, Dung Nguyen, Hau Lam, Phong Nguyen, Jennifer Fowler, Minh-Triet Tran, Ngan Le; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1981-1990

The timing of cell divisions in early embryos during the In-Vitro Fertilization (IVF) process is a key predictor of embryo viability. However, observing cell divisions in Time-Lapse Monitoring (TLM) is a time-consuming process and highly depends on experts. In this paper, we propose EmbryosFormer, a computational model to automatically detect and classify cell divisions from original time-lapse images. Our proposed network is designed as an encoder-decoder deformable transformer with collaborative heads. The transformer contracting path predicts per-image labels and is optimized by a classification head. The transformer expanding path models the temporal coherency between embryo images to ensure monotonic non-decreasing constraint and is optimized by a segmentation head. Both contracting and expanding paths are synergetically learned by a collaboration head. We have benchmarked our proposed EmbryosFormer on two datasets: a public dataset with mouse embryos with 8-cell stage and an in-house dataset with human embryos with 4-cell stage. Source code: https://github.com/UARK-AICV/Embryos.

**********

Hyperspherical Quantization: Toward Smaller and More Accurate Models

Dan Liu, Xi Chen, Chen Ma, Xue Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5262-5272

Model quantization enables the deployment of deep neural networks under resource-constrained devices. Vector quantization aims at reducing the model size by indexing model weights with full-precision embeddings, i.e., codewords, while the index needs to be restored to 32-bit during computation. Binary and other low-precision quantization methods can reduce the model size up to 32x, however, at the cost of a considerable accuracy drop. In this paper, we propose an efficient framework for ternary quantization to produce smaller and more accurate compressed models. By integrating hyperspherical learning, pruning and reinitialization, our proposed Hyperspherical Quantization (HQ) method reduces the cosine distance between the full-precision and ternary weights, thus reducing the bias of the straight-through gradient estimator during ternary quantization. Compared with existing work at similar compression levels ( 30x,  40x), our method significantly improves the test accuracy and reduces the model size.

**********

TinyHD: Efficient Video Saliency Prediction With Heterogeneous Decoders Using Hierarchical Maps Distillation

Feiyan Hu, Simone Palazzo, Federica Proietto Salanitri, Giovanni Bellitto, Morteza Moradi, Concetto Spampinato, Kevin McGuinness; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2051-2060

Video saliency prediction has recently attracted attention of the research community, as it is an upstream task for several practical applications. However, current solutions are particurly computationally demanding, especially due to the wide usage of spatio-temporal 3D convolutions. We observe that, while different model architectures achieve similar performance on benchmarks, visual variations between predicted saliency maps are still significant. Inspired by this intuition, we propose a lightweight model that employs multiple simple heterogeneous decoders and adopts several practical approaches to improve accuracy while keeping

computational costs low, such as hierarchical multi-map knowledge distillation, multi-output saliency prediction, unlabeled auxiliary datasets and channel reduction with teacher assistant supervision. Our approach achieves saliency prediction accuracy on par or better than state-of-the-art methods on DFH1K, UCF-Sports and Hollywood2 benchmarks, while enhancing significantly the efficiency of the model.

**********

Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, Bastian Wandt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2592-2602

Industrial defect detection is commonly addressed with anomaly detection (AD) methods where no or only incomplete data of potentially occurring defects is available. This work discovers previously unknown problems of student-teacher approaches for AD and proposes a solution, where two neural networks are trained to produce the same output for the defect-free training examples. The core assumption of student-teacher networks is that the distance between the outputs of both networks is larger for anomalies since they are absent in training. However, previous methods suffer from the similarity of student and teacher architecture, such that the distance is undesirably small for anomalies. For this reason, we propose asymmetric student-teacher networks (AST). We train a normalizing flow for density estimation as a teacher and a conventional feed-forward network as a student to trigger large distances for anomalies: The bijectivity of the normalizing flow enforces a divergence of teacher outputs for anomalies compared to normal data. Outside the training distribution the student cannot imitate this divergence due to its fundamentally different architecture. Our AST network compensates for wrongly estimated likelihoods by a normalizing flow, which was alternatively used for anomaly detection in previous work. We show that our method produces state-of-the-art results on the two currently most relevant defect detection datasets MVTec AD and MVTec 3D-AD regarding image-level anomaly detection on RGB and 3D data.

**********

Calibrating Deep Neural Networks Using Explicit Regularisation and Dynamic Data Pruning

Rishabh Patra, Ramya Hebbalaguppe, Tirtharaj Dash, Gautam Shroff, Lovekesh Vig; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1541-1549

Deep neural networks are prone to miscalibrated predictions, often exhibiting a mismatch between the target output and the generated sample confidence scores. Contemporary model calibration techniques mitigate the problem of overconfident predictions by pushing down the confidence of the winning class while increasing the confidence of the remaining classes across all test samples. However, from a deployment perspective, an ideal model would (i) generate well-calibrated predictions for high-confidence samples (say, Prob > 0.95) and (ii) generate a higher proportion of legitimate high-confidence samples. To this end, we propose a novel regularization technique that can be used with classification losses, leading to state-of-the-art calibrated predictions at test time; From a deployment standpoint in safety-critical applications, only high-confidence samples from a well-calibrated model are of interest, as the remaining samples have to undergo manual inspection. Predictive confidence reduction of these potentially "high-confidence samples" is a downside of existing calibration approaches. To mitigate this, we propose a dynamic train-time data pruning strategy which prunes low confidence samples every few epochs, providing an increase in confident yet calibrated samples. We demonstrate state-of-the-art calibration performance across image classification benchmarks, reducing training time without much compromise in accuracy. We provide insights into our dynamic pruning strategy showing that pruning low-confidence training samples lead to an increase in high-confidence samples at test time.

**********

Searching Efficient Neural Architecture With Multi-Resolution Fusion Transformer

for Appearance-Based Gaze Estimation
Vikrant Nagpure, Kenji Okuma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 890-899
For aiming at a more accurate appearance-based gaze estimation, a series of recent works propose to use transformers or high-resolution networks in several ways which achieve state-of-the-art, but such works lack efficiency for real-time applications on edge computing devices. In this paper, we propose a compact model to precisely and efficiently solve gaze estimation. The proposed model includes 1) a Neural Architecture Search(NAS)-based multi-resolution feature extractor for extracting feature maps with global and local information which are essential for this task and 2) a novel multi-resolution fusion transformer as the gaze estimation head for efficiently estimating gaze values by fusing the extracted feature maps. We search our proposed model, called GazeNAS-ETH, on the ETH-XGaze dataset. We confirmed through experiments that GazeNAS-ETH achieved state-of-the-art on Gaze360, MPIIFaceGaze, RTGENE, and EYEDIAP datasets, while having only about 1M parameters and using only 0.28 GFLOPs, which is significantly less compared to previous state-of-the-art models, making it easier to deploy for real-time applications.
**********

Stop or Forward: Dynamic Layer Skipping for Efficient Action Recognition
Jonghyeon Seon, Jaedong Hwang, Jonghwan Mun, Bohyung Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3361-3370
One of the challenges for analyzing video contents (e.g., actions) is high computational cost, especially for the tasks that require processing densely sampled frames in a long video. We present a novel efficient action recognition algorithm, which allocates computational resources adaptively to individual frames depending on their relevance and significance. Specifically, our algorithm adopts LSTM-based policy modules and sequentially estimates the usefulness of each frame based on their intermediate representations. If a certain frame is unlikely to be helpful for recognizing actions, our model stops forwarding the features to the rest of the layers and starts to consider the next sampled frame. We further reduce the computational cost of our approach by introducing a simple yet effective early termination strategy during the inference procedure. We evaluate the proposed algorithm on three public benchmarks: ActivityNet-v1.3, Mini-Kinetics, and THUMOS'14. Our experiments show that the proposed approach achieves outstanding trade-off between accuracy and efficiency in action recognition.
**********

Overlap-Guided Gaussian Mixture Models for Point Cloud Registration
Guofeng Mei, Fabio Poiesi, Cristiano Saltori, Jian Zhang, Elisa Ricci, Nicu Sebe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4511-4520
Probabilistic 3D point cloud registration methods have shown competitive performance in overcoming noise, outliers, and density variations. However, registering point cloud pairs in the case of partial overlap is still a challenge. This paper proposes a novel overlap-guided probabilistic registration approach that computes the optimal transformation from matched Gaussian Mixture Model (GMM) parameters. We reformulate the registration problem as the problem of aligning two Gaussian mixtures such that a statistical discrepancy measure between the two corresponding mixtures is minimized. We introduce a Transformer-based detection module to detect overlapping regions and represent the input point clouds using GMMs by guiding their alignment through overlap scores computed by this detection module. Experiments show that our method achieves superior registration accuracy and efficiency than state-of-the-art methods when handling point clouds with partial overlap and different densities on synthetic and real-world datasets.
**********

Magnification Prior: A Self-Supervised Method for Learning Representations on Breast Cancer Histopathological Images
Prakash Chandra Chhipa, Richa Upadhyay, Gustav Grund Pihlgren, Rajkumar Saini, Seiichi Uchida, Marcus Liwicki; Proceedings of the IEEE/CVF Winter Conference on

Applications of Computer Vision (WACV), 2023, pp. 2717-2727
This work presents a novel self-supervised pre-training method to learn efficient representations without labels on histopathology medical images utilizing magnification factors. Other state-of-the-art works mainly focus on fully supervised learning approaches that rely heavily on human annotations. However, the scarcity of labeled and unlabeled data is a long-standing challenge in histopathology. Currently, representation learning without labels remains unexplored in the histopathology domain. The proposed method, Magnification Prior Contrastive Similarity (MPCS), enables self-supervised learning of representations without labels on small-scale breast cancer dataset BreakHis by exploiting magnification factor, inductive transfer, and reducing human prior. The proposed method matches fully supervised learning state-of-the-art performance in malignancy classification when only 20% of labels are used in fine-tuning and outperform previous works in fully supervised learning settings for three public breast cancer datasets, including BreakHis. Further, It provides initial support for a hypothesis that reducing human-prior leads to efficient representation learning in self-supervision, which will need further investigation. The implementation of this work is available online on GitHub.
**********

Semi-Supervised Learning for Sparsely-Labeled Sequential Data: Application to Healthcare Video Processing
Florian Dubost, Erin Hong, Siyi Tang, Nandita Bhaskhar, Christopher Lee-Messer, Daniel Rubin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1890-1899
Labeled data is a critical resource for training and evaluating machine learning models. However, many real-life datasets are only partially labeled. We propose a semi-supervised machine learning training strategy to improve event detection performance on sequential data, such as video recordings, when only sparse labels are available, such as event start times without their corresponding end times. Our method uses noisy guesses of the events' end times to train event detection models. Depending on how conservative these guesses are, mislabeled samples may be introduced into the training set. We further propose a mathematical model for explaining and estimating the evolution of the classification performance for increasingly noisier end time estimates. We show that neural networks can improve their detection performance by leveraging more training data with less conservative approximations despite the higher proportion of incorrect labels. We adapt sequential versions of CIFAR-10 and MNIST, and use the Berkeley MHAD and HMBD51 video datasets to empirically evaluate our method, and find that our risk-tolerant strategy outperforms conservative estimates by 3.5 points of mean average precision for CIFAR, 30 points for MNIST, 3 points for MHAD, and 14 points for HMBD51. Then, we leverage the proposed training strategy to tackle a real-life application: processing continuous video recordings of epilepsy patients, and show that our method outperforms baseline labeling methods by 17 points of average precision, and reaches a classification performance similar to that of fully supervised models.
**********

WHFL: Wavelet-Domain High Frequency Loss for Sketch-to-Image Translation
Min Woo Kim, Nam Ik Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 744-754
Even a rough sketch can effectively convey the descriptions of objects, as humans can imagine the original shape from the sketch. The sketch-to-photo translation is a computer vision task that enables a machine to do this imagination, taking a binary sketch image and generating plausible RGB images corresponding to the sketch. Hence, deep neural networks for this task should learn to generate a wide range of frequencies because most parts of the input (binary sketch image) are composed of DC signals. In this paper, we propose a new loss function named Wavelet-domain High-Frequency Loss (WHFL) to overcome the limitations of previous methods that tend to have a bias toward low frequencies. The proposed method emphasizes the loss on the high frequencies by designing a new weight matrix imposing larger weights on the high bands. Unlike existing hand-craft methods that con

trol frequency weights using binary masks, we use the matrix with finely control led elements according to frequency scales. The WHFL is designed in a multi-scal e form, which lets the loss function focus more on the high frequency according to decomposition levels. We use the WHFL as a complementary loss in addition to conventional ones defined in the spatial domain. Experiments show we can improve the qualitative and quantitative results in both spatial and frequency domains. Additionally, we attempt to verify the WHFL's high-frequency generation capabil ity by defining a new evaluation metric named Unsigned Euclidean Distance Field Error (UEDFE).
**********

Enabling ISPless Low-Power Computer Vision
Gourav Datta, Zeyu Liu, Zihan Yin, Linyu Sun, Akhilesh R. Jaiswal, Peter A. Beer el; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi sion (WACV), 2023, pp. 2430-2439
Current computer vision (CV) systems use an image signal processing (ISP) unit t o convert the high resolution raw images captured by image sensors to visually p leasing RGB images. Typically, CV models are trained on these RGB images and hav e yielded state-of-the-art (SOTA) performance on a wide range of complex vision tasks, such as object detection. In addition, in order to deploy these models on resource-constrained low-power devices, recent works have proposed in-sensor an d in-pixel computing approaches that try to partly/fully bypass the ISP and yiel d significant bandwidth reduction between the image sensor and the CV processing unit by downsampling the activation maps in the initial convolutional neural ne twork (CNN) layers. However, direct inference on the raw images degrades the tes t accuracy due to the difference in covariance of the raw images captured by the image sensors compared to the ISP-processed images used for training. Moreover, it is difficult to train deep CV models on raw images, because most (if not all ) large-scale open-source datasets consist of RGB images. To mitigate this conce rn, we propose to invert the ISP pipeline, which can convert the RGB images of a ny dataset to its raw counterparts, and enable model training on raw images. We release the raw version of the COCO dataset, a large-scale benchmark for generic high-level vision tasks. For ISP-less CV systems, training on these raw images result in a 7.1% increase in test accuracy on the visual wake works (VWW) datase t compared to relying on training with traditional ISP-processed RGB datasets. T o further improve the accuracy of ISP-less CV models and to increase the energy and bandwidth benefits obtained by in-sensor/in-pixel computing, we propose an e nergy-efficient form of analog in-pixel demosaicing that may be coupled with in- pixel CNN computations. When evaluated on raw images captured by real sensors fr om the PASCALRAW dataset, our approach results in a 8.1% increase in mAP.
**********

VSGD-Net: Virtual Staining Guided Melanocyte Detection on Histopathological Imag es
Kechun Liu, Beibin Li, Wenjun Wu, Caitlin May, Oliver Chang, Stevan Knezevich, L isa Reisch, Joann Elmore, Linda Shapiro; Proceedings of the IEEE/CVF Winter Conf erence on Applications of Computer Vision (WACV), 2023, pp. 1918-1927
Detection of melanocytes serves as a critical prerequisite in assessing melanocy tic growth patterns when diagnosing melanoma and its precursor lesions on skin b iopsy specimens. However, this detection is challenging due to the visual simila rity of melanocytes to other cells in routine Hematoxylin and Eosin (H&E) staine d images, leading to the failure of current nuclei detection methods. Stains suc h as Sox10 can mark melanocytes, but they require an additional step and expense and thus are not regularly used in clinical practice. To address these limitati ons, we introduce VSGD-Net, a novel detection network that learns melanocyte ide ntification through virtual staining from H&E to Sox10. The method takes only ro utine H&E images during inference, resulting in a promising approach to support pathologists in the diagnosis of melanoma. To the best of our knowledge, this is the first study that investigates the detection problem using image synthesis f eatures between two distinct pathology stainings. Extensive experimental results show that our proposed model outperforms state-of-the-art nuclei detection meth ods.

**********
CG-NeRF: Conditional Generative Neural Radiance Fields for 3D-Aware Image Synthesis
Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, Jaegul Choo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 724-733
Recent generative models based on neural radiance fields (NeRF) achieve the generation of diverse 3D-aware images. Despite the success, their applicability can be further expanded by incorporating with various types of user-specified conditions such as text and images. In this paper, we propose a novel approach called the conditional generative neural radiance fields (CG-NeRF), which generates multi-view images that reflect multimodal input conditions such as images or text. However, generating 3D-aware images from multimodal conditions bears several challenges. First, each condition type has different amount of information - e.g., the amount of information in text and color images are significantly different. Furthermore, the pose-consistency is often violated when diversifying the generated images from input conditions. Addressing such challenges, we propose 1) a unified architecture that effectively handles multiple types of conditions, and 2) the pose-consistent diversity loss for generating various images while maintaining the view consistency. Experimental results show that the proposed method maintains consistent image quality on various multimodal condition types and achieves superior fidelity and diversity compared to the existing NeRF-based generative models.
**********
A Priority Map for Vision-and-Language Navigation With Trajectory Plans and Feature-Location Cues
Jason Armitage, Leonardo Impett, Rico Sennrich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1094-1103
In a busy city street, a pedestrian surrounded by distractions can pick out a single sign if it is relevant to their route. Artificial agents in outdoor Vision-and-Language Navigation (VLN) are also confronted with detecting supervisory signal on environment features and location in inputs. To boost the prominence of relevant features in transformer-based systems without costly preprocessing and pretraining, we take inspiration from priority maps - a mechanism described in neuropsychological studies. We implement a novel priority map module and pretrain on auxiliary tasks using low-sample datasets with high-level representations of routes and environment-related references to urban features. A hierarchical process of trajectory planning - with subsequent parameterised visual boost filtering on visual inputs and prediction of corresponding textual spans - addresses the core challenge of cross-modal alignment and feature-level localisation. The priority map module is integrated into a feature-location framework that doubles the task completion rates of standalone transformers and attains state-of-the-art performance for transformer-based systems on the Touchdown benchmark for VLN. We release code (https://github.com/JasonArmitage-res/PM-VLN) and data (https://zenodo.org/record/6891965#.YtwoS3ZBxD8).
**********
A Quality Aware Sample-to-Sample Comparison for Face Recognition
Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan, Ali Zafari, Moktari Mostofa, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6129-6138
Currently available face datasets mainly consist of a large number of high-quality and a small number of low-quality samples. As a result, a Face Recognition (FR) network fails to learn the distribution of low-quality samples since they are less frequent during training (underrepresented). Moreover, current state-of-the-art FR training paradigms are based on the sample-to-center comparison (i.e., Softmax-based classifier), which results in a lack of uniformity between train and test metrics. This work integrates a quality-aware learning process at the sample level into the classification training paradigm (QAFace). In this regard, Softmax centers are adaptively guided to pay more attention to low-quality samples by using a quality-aware function. Accordingly, QAFace adds a quality-based ad

justment to the updating procedure of the Softmax-based classifier to improve th
e performance on the underrepresented low-quality samples. Our method adaptively
 finds and assigns more attention to the recognizable low-quality samples in the
 training datasets. In addition, QAFace ignores the unrecognizable low-quality s
amples using the feature magnitude as a proxy for quality. As a result, QAFace p
revents class centers from getting distracted from the optimal direction. The pr
oposed method is superior to the state-of-the-art algorithms in extensive experi
mental results on the CFP-FP, LFW, CPLFW, CALFW, AgeDB, IJB-B, and IJB-C dataset
s.
**********

Computer Vision to the Rescue: Infant Postural Symmetry Estimation From Incongru
ent Annotations
Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, Sarah Ostadabbas; Proce
edings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC
V), 2023, pp. 1909-1917
Bilateral postural symmetry plays a key role as a potential risk marker for auti
sm spectrum disorder (ASD) and as a symptom of congenital muscular torticollis (
CMT) in infants, but current methods of assessing symmetry require laborious cli
nical expert assessments. In this paper, we develop a computer vision based infa
nt symmetry assessment system, leveraging 3D human pose estimation for infants.
Evaluation and calibration of our system against ground truth assessments is com
plicated by our findings from a survey of human ratings of angle and symmetry, t
hat such ratings exhibit low inter-rater reliability. To rectify this, we develo
p a Bayesian estimator of the ground truth derived from a probabilistic graphica
l model of fallible human raters. We show that the 3D infant pose estimation mod
el can achieve 68% area under the receiver operating characteristic curve perfor
mance in predicting the Bayesian aggregate labels, compared to only 61% from a 2
D infant pose estimation model and 60% from a 3D adult pose estimation model, hi
ghlighting the importance of 3D poses and infant domain knowledge in assessing i
nfant body symmetry. Our survey analysis also suggests that human ratings are su
sceptible to higher levels of bias and inconsistency, and hence our final 3D pos
e-based symmetry assessment system is calibrated but not directly supervised by
Bayesian aggregate human ratings, yielding higher levels of consistency and lowe
r levels of inter-limb assessment bias.
**********

Recovering Fine Details for Neural Implicit Surface Reconstruction
Decai Chen, Peng Zhang, Ingo Feldmann, Oliver Schreer, Peter Eisert; Proceedings
 of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20
23, pp. 4330-4339
Recent works on implicit neural representations have made significant strides. L
earning implicit neural surfaces using volume rendering has gained popularity in
 multi-view reconstruction without 3D supervision. However, accurately recoverin
g fine details is still challenging, due to the underlying ambiguity of geometry
 and appearance representation. In this paper, we present D-NeuS, a volume rende
ring-base neural implicit surface reconstruction method capable to recover fine
geometry details, which extends NeuS by two additional loss functions targeting
enhanced reconstruction quality. First, we encourage the rendered surface points
 from alpha compositing to have zero signed distance values, alleviating the geo
metry bias arising from transforming SDF to density for volume rendering. Second
, we impose multi-view feature consistency on the surface points, derived by int
erpolating SDF zero-crossings from sampled points along rays. Extensive quantita
tive and qualitative results demonstrate that our method reconstructs high-accur
acy surfaces with details, and outperforms the state of the art.
**********

More Knowledge, Less Bias: Unbiasing Scene Graph Generation With Explicit Ontolo
gical Adjustment
Zhanwen Chen, Saed Rezayi, Sheng Li; Proceedings of the IEEE/CVF Winter Conferen
ce on Applications of Computer Vision (WACV), 2023, pp. 4023-4032
Scene graph generation (SGG) models seek to detect relationships between objects
 in a given image. One challenge in this area is the biased distribution of pred

icates in the dataset and the semantic space. Recent works incorporating knowled ge graphs with scene graphs prove effective in improving recall for the tail pre dicate classes. Moreover, many recent SGG approaches with promising results expl icitly redistribute the predicates in both the training process and in the predi ction step. To incorporate external knowledge, we construct a commonsense knowle dge graph by integrating ConceptNet and Wikidata. To explicitly unbias SGG with knowledge in the reasoning process, we propose a novel framework, Explicit Ontol ogical Adjustment (EOA), to adjust the graph model predictions with knowledge pr iors. We use the edge matrix from the commonsense knowledge graph as a module in the graph neural network model to refine the relationship detection process. Th is module proves effective in alleviating the long-tail distribution of predicat es. When combined, we show that these modules achieve state-of-the-art performan ce on the Visual Genome dataset in most cases. The source code will be made publ icly available.

**********

Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation

Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, Jürgen Beyerer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visio n (WACV), 2023, pp. 6068-6077

Although human action anticipation is a task which is inherently multi-modal, st ate-of-the-art methods on well known action anticipation datasets leverage this data by applying ensemble methods and averaging scores of uni-modal anticipation networks. In this work we introduce transformer based modality fusion technique s, which unify multi-modal data at an early stage. Our Anticipative Feature Fusi on Transformer (AFFT) proves to be superior to popular score fusion approaches a nd presents state-of-the-art results outperforming previous methods on EpicKitch ens-100 and EGTEA Gaze+. Our model is easily extensible and allows for adding ne w modalities without architectural changes. Consequently, we extracted audio fea tures on EpicKitchens-100 which we add to the set of commonly used features in t he community.

**********

Text-Guided Object Detector for Multi-Modal Video Question Answering

Ruoyue Shen, Nakamasa Inoue, Koichi Shinoda; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1032-1042

Video Question Answering (Video QA) is a task to answer a text-format question b ased on the understanding of linguistic semantics, visual information, and also linguistic-visual alignment in the video. In Video QA, an object detector pre-tr ained with large-scale datasets, such as Faster R-CNN, has been widely used to e xtract visual representations from video frames. However, it is not always able to precisely detect the objects needed to answer the question because of the dom ain gaps between the datasets for training the object detector and those for Vid eo QA. In this paper, we propose a text-guided object detector (TGOD), which tak es text question-answer pairs and video frames as inputs, detects the objects re levant to the given text, and thus provides intuitive visualization and interpre table results. Our experiments using the STAGE framework on the TVQA+ dataset sh ow the effectiveness of our proposed detector. It achieves a 2.02 points improve ment in accuracy of QA, 12.13 points improvement in object detection (mAP50), 1. 1 points improvement in temporal location, and 2.52 points improvement in ASA ov er the STAGE original detector.

**********

Multi-Scale Contrastive Learning for Complex Scene Generation

Hanbit Lee, Youna Kim, Sang-goo Lee; Proceedings of the IEEE/CVF Winter Conferen ce on Applications of Computer Vision (WACV), 2023, pp. 764-774

Recent advances in Generative Adversarial Networks (GANs) have enabled photo-rea listic synthesis of single object images. Yet, modeling more complex distributio ns, such as scenes with multiple objects, remains challenging. The difficulty st ems from the incalculable variety of scene configurations which contain multiple objects of different categories placed at various locations. In this paper, we aim to alleviate the difficulty by enhancing the discriminative ability of the d iscriminator through a locally defined self-supervised pretext task. To this end

, we design a discriminator to leverage multi-scale local feedback that guides t
he generator to better model local semantic structures in the scene. Then, we re
quire the discriminator to carry out pixel-level contrastive learning at multipl
e scales to enhance discriminative capability on local regions. Experimental res
ults on several challenging scene datasets show that our method improves the syn
thesis quality by a substantial margin compared to state-of-the-art baselines.
**********

Beyond RGB: Scene-Property Synthesis With Neural Radiance Fields
Mingtong Zhang, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Yu-Xiong Wang; Proce
edings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC
V), 2023, pp. 795-805
Comprehensive 3D scene understanding, both geometrically and semantically, is im
portant for real-world applications such as robot perception. Most of the existi
ng work has focused on developing data-driven discriminative models for scene un
derstanding. This paper provides a new approach to scene understanding, from a s
ynthesis model perspective, by leveraging the recent progress on implicit scene
representation and neural rendering. Building upon the great success of Neural R
adiance Fields (NeRFs), we introduce Scene-Property Synthesis with NeRF (SS-NeRF
) that is able to not only render photo-realistic RGB images from novel viewpoin
ts, but also render various accurate scene properties (e.g., appearance, geometr
y, and semantics). By doing so, we facilitate addressing a variety of scene unde
rstanding tasks under a unified framework, including semantic segmentation, surf
ace normal estimation, reshading, keypoint detection, and edge detection. Our SS
-NeRF framework can be a powerful tool for bridging generative learning and disc
riminative learning, and thus be beneficial to the investigation of a wide range
 of interesting problems, such as studying task relationships within a synthesis
 paradigm, transferring knowledge to novel tasks, facilitating downstream discri
minative tasks as ways of data augmentation, and serving as auto-labeller for da
ta creation. Our code is available at https://github.com/zsh2000/SS-NeRF.
**********

Closer Look at the Transferability of Adversarial Examples: How They Fool Differ
ent Models Differently
Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H. Nguyen, Isao Echizen; Proc
eedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WA
CV), 2023, pp. 1360-1368
Deep neural networks are vulnerable to adversarial examples (AEs), which have ad
versarial transferability: AEs generated for the source model can mislead anothe
r (target) model's predictions. However, the transferability has not been unders
tood in terms of to which class target model's predictions were misled (i.e., cl
ass-aware transferability). In this paper, we differentiate the cases in which a
 target model predicts the same wrong class as the source model ("same mistake")
 or a different wrong class ("different mistake") to analyze and provide an expl
anation of the mechanism. We find that (1) AEs tend to cause same mistakes, whic
h correlates with "non-targeted transferability"; however, (2) different mistake
s occur even between similar models, regardless of the perturbation size. Furthe
rmore, we present evidence that the difference between same mistakes and differe
nt mistakes can be explained by non-robust features, predictive but human-uninte
rpretable patterns: different mistakes occur when non-robust features in AEs are
 used differently by models. Non-robust features can thus provide consistent exp
lanations for the class-aware transferability of AEs.
**********

TTTFlow: Unsupervised Test-Time Training With Normalizing Flow
David Osowiechi, Gustavo A. Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani,
Ismail Ben Ayed, Christian Desrosiers; Proceedings of the IEEE/CVF Winter Confer
ence on Applications of Computer Vision (WACV), 2023, pp. 2126-2134
A major problem of deep neural networks for image classification is their vulner
ability to domain changes at test-time. Recent methods have proposed to address
this problem with test-time training (TTT), where a two-branch model is trained
to learn a main classification task and also a self-supervised task used to perf
orm test-time adaptation. However, these techniques require defining a proxy tas

k specific to the target application. To tackle this limitation, we propose TTTF low: a Y-shaped architecture using an unsupervised head based on Normalizing Flo ws to learn the normal distribution of latent features and detect domain shifts in test examples. At inference, keeping the unsupervised head fixed, we adapt th e model to domain-shifted examples by maximizing the log likelihood of the Norma lizing Flow. Our results show that our method can significantly improve the accu racy with respect to previous works.
**********

Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors
Tobias Riedlinger, Matthias Rottmann, Marius Schubert, Hanno Gottschalk; Proceed ings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) , 2023, pp. 3921-3931
The majority of uncertainty quantification methods for deep object detectors are based on the network output, such as sampling strategies like Monte-Carlo dropo ut or deep ensembles with straight-forward transfers to object detection. Here, we study gradient-based uncertainty features for object detection. We show that they contain information orthogonal to that of common, output-based uncertainty approximation methods. Meta classification and meta regression are used to produ ce confidence estimates using gradient features and other methods which are appl icable to numerous object detection architectures. Our results show that gradien t uncertainty itself performs on par with stateof-the-art methods across differe nt detectors and datasets. We find that combined meta classifiers outperform sta ndalone models. This suggests that sampling strategies may be supplemented by gr adient-based uncertainty to obtain improved confidences, contributing to the pro babilistic reliability of object detectors in down-stream applications.
**********

M-FUSE: Multi-Frame Fusion for Scene Flow Estimation
Lukas Mehl, Azin Jahedi, Jenny Schmalfuss, Andrés Bruhn; Proceedings of the IEEE /CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2020 -2029
Recently, neural network for scene flow estimation show impressive results on au tomotive data such as the KITTI benchmark. However, despite of using sophisticat ed rigidity assumptions and parametrizations, such networks are typically limite d to only two frame pairs which does not allow them to exploit temporal informat ion. In our paper we address this shortcoming by proposing a novel multi-frame a pproach that considers an additional preceding stereo pair. To this end, we proc eed in two steps: Firstly, building upon the recent RAFT-3D approach, we develop an improved two-frame baseline by incorporating an advanced stereo method. Seco ndly, and even more importantly, exploiting the specific modeling concepts of RA FT-3D, we propose a U-Net architecture that performs a fusion of forward and bac kward flow estimates and hence allows to integrate temporal information on deman d. Experiments on the KITTI benchmark do not only show that the advantages of th e improved baseline and the temporal fusion approach complement each other, they also demonstrate that the computed scene flow is highly accurate. More precisel y, our approach ranks second overall and first for the even more challenging for eground objects, in total outperforming the original RAFT-3D method by more than 16%. Code is available at https://github.com/cv-stuttgart/M-FUSE.
**********

Unifying Distribution Alignment as a Loss for Imbalanced Semi-Supervised Learnin g
Justin Lazarow, Kihyuk Sohn, Chen-Yu Lee, Chun-Liang Li, Zizhao Zhang, Tomas Pfi ster; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5644-5653
While remarkable progress has been made in imbalanced supervised learning, less attention has been given to the setting of imbalanced semi-supervised learning ( SSL) where not only is few labeled data provided, but the underlying data distri bution can be severely imbalanced. Recent work requires both complicated samplin g strategies of pseudo-labeled unlabeled data and distribution alignment of the pseudo-label distribution to accommodate this imbalance. We present a novel appr oach that relies only on a form of a distribution alignment but no sampling stra

tegy where rather than aligning the pseudo-labels during inference, we move the distribution alignment component into the respective cross entropy loss computations for both the supervised and unsupervised losses. This alignment compensates for both imbalance in the data and the eventual distributional shift present during evaluation. Altogether, this provides a unified strategy that offers both significantly reduced training requirements and improved performance across both low and richly labeled regimes and over varying degrees of imbalance. In experiments, we validate the efficacy of our method on SSL variants of CIFAR10-LT, CIFAR100-LT, and ImageNet-127. On ImageNet-127, our method shows 1.6% accuracy improvement over CReST with an 80% training time reduction and is competitive with other SOTA methods.

**********

Real-Time Restoration of Dark Stereo Images

Mohit Lamba, M. V. A. Suhas Kumar, Kaushik Mitra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4914-4924

Low-light image enhancement has been an actively researched area for decades and has produced excellent night-time single-image, video, and Light Field restoration methods. Despite these comprehensive studies, the problem of extreme low-light stereo image enhancement has been mostly ignored. Addressing this problem can enable night-time capabilities to several applications such as smartphones and self-driving cars. We propose a light-weight and fast hybrid U-net architecture for low-light stereo image enhancement. In the initial few scale spaces, we process the left and right features individually, because the two features do not align well due to large disparity. At coarser scale-spaces, the disparity between left and right features decreases and the network's receptive field increases. We use this fact to reduce computations by simultaneously processing the left and right features, which also benefits epipole preservation. As our architecture does not use any 3D convolution for fast inference, we use an Epipole-Aware loss module to train our network. This module computes quick and coarse depth estimates to better enforce the epipolar constraints. Extensive benchmarking in terms of visual enhancement and downstream depth estimation shows that our architecture not only performs significantly better but also offers 4-60 xspeed-up with 15-100 xlower floating point operations, suitable for real-world deployment.

**********

Automatically Annotating Indoor Images With CAD Models via RGB-D Scans

Stefan Ainetter, Sinisa Stekovic, Friedrich Fraundorfer, Vincent Lepetit; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3156-3164

We present an automatic method for annotating images of indoor scenes with the CAD models of the objects by relying on RGB-D scans. Through a visual evaluation by 3D experts, we show that our method retrieves annotations that are at least as accurate as manual annotations, and can thus be used as ground truth without the burden of manually annotating 3D data. We do this using an analysis-by-synthesis approach, which compares renderings of the CAD models with the captured scene. We introduce a 'cloning procedure' that identifies objects that have the same geometry, to annotate these objects with the same CAD models. This allows us to obtain complete annotations for the ScanNet dataset and the recent ARKitScenes dataset. We will release these annotations publicly, as we believe they will be very useful for the computer vision community.

**********

CFL-Net: Image Forgery Localization Using Contrastive Learning

Fahim Faisal Niloy, Kishor Kumar Bhaumik, Simon S. Woo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4642-4651

Conventional forgery localizing methods usually rely on different forgery footprints such as JPEG artifacts, edge inconsistency, camera noise, etc., with cross-entropy loss to locate manipulated regions. However, these methods have the disadvantage of over-fitting and focusing on only a few specific forgery footprints. On the other hand, real-life manipulated images are generated via a wide variety of forgery operations and thus, leave behind a wide variety of forgery footpri

nts. Therefore, we need a more general approach for image forgery localization t hat can work well on a variety of forgery conditions. A key assumption in underl ying forged region localization is that there remains a difference of feature di stribution between untampered and manipulated regions in each forged image sampl e, irrespective of the forgery type. In this paper, we aim to leverage this diff erence of feature distribution to aid in image forgery localization. Specificall y, we use contrastive loss to learn mapping into a feature space where the featu res between untampered and manipulated regions are well-separated for each image . Also, our method has the advantage of localizing manipulated region without re quiring any prior knowledge or assumption about the forgery type. We demonstrate that our work outperforms several existing methods on three benchmark image man ipulation datasets. Code is available at https://github.com/niloy193/CFLNet.
**********

The CropAndWeed Dataset: A Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation

Daniel Steininger, Andreas Trondl, Gerardus Croonen, Julia Simon, Verena Widhalm ; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visi on (WACV), 2023, pp. 3729-3738

Precision Agriculture and especially the application of automated weed intervent ion represents an increasingly essential research area, as sustainability and ef ficiency considerations are becoming more and more relevant. While the potential s of Convolutional Neural Networks for detection, classification and segmentatio n tasks have successfully been demonstrated in other application areas, this rel atively new field currently lacks the required quantity and quality of training data for such a highly data-driven approach. Therefore, we propose a novel large -scale image dataset specializing in the fine-grained identification of 74 relev ant crop and weed species with a strong emphasis on data variability. We provide annotations of labeled bounding boxes, semantic masks and stem positions for ab out 112k instances in more than 8k high-resolution images of both real-world agr icultural sites and specifically cultivated outdoor plots of rare weed types. Ad ditionally, each sample is enriched with an extensive set of meta-annotations re garding environmental conditions and recording parameters. We furthermore conduc t benchmark experiments for multiple learning tasks on different variants of the dataset to demonstrate its versatility and provide examples of useful mapping s chemes for tailoring the annotated data to the requirements of specific applicat ions. In the course of the evaluation, we furthermore demonstrate how incorporat ing multiple species of weeds into the learning process increases the accuracy o f crop detection. Overall, the evaluation clearly demonstrates that our dataset represents an essential step towards overcoming the data gap and promoting furth er research in the area of Precision Agriculture.
**********

Ego-Vehicle Action Recognition Based on Semi-Supervised Contrastive Learning

Chihiro Noguchi, Toshihiro Tanizawa; Proceedings of the IEEE/CVF Winter Conferen ce on Applications of Computer Vision (WACV), 2023, pp. 5988-5998

In recent years, many automobiles have been equipped with cameras, which have ac cumulated an enormous amount of video footage of driving scenes. Autonomous driv ing demands the highest level of safety, for which even unimaginably rare drivin g scenes have to be collected in training data to improve the recognition accura cy for specific scenes. However, it is prohibitively costly to find very few spe cific scenes from an enormous amount of videos. In this article, we show that pr oper video-to-video distances can be defined by focusing on ego-vehicle actions. It is well known that existing methods based on supervised learning cannot hand le videos that do not fall into predefined classes, though they work well in def ining video-to-video distances in the embedding space between labeled videos. To tackle this problem, we propose a method based on semi-supervised contrastive l earning. We consider two related but distinct contrastive learning: standard gra ph contrastive learning and our proposed SOIA-based contrastive learning. We obs erve that the latter approach can provide more sensible video-to-video distances between unlabeled videos. Next, the effectiveness of our method is quantified b y evaluating the classification performance of the ego-vehicle action recognitio

n using HDD dataset, which shows that our method including unlabeled data in tra
ining significantly outperforms the existing methods using only labeled data in
training.
**********

Frame Interpolation for Dynamic Scenes With Implicit Flow Encoding
Pedro Figueirêdo, Avinash Paliwal, Nima Khademi Kalantari; Proceedings of the IE
EE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 21
8-228
In this paper, we propose an algorithm to interpolate between a pair of images o
f a dynamic scene. While in the past years significant progress in frame interpo
lation has been made, current approaches are not able to handle images with brig
htness and illumination changes, which are common even when the images are captu
red shortly apart. We propose to address this problem by taking advantage of the
 existing optical flow methods that are highly robust to the variations in the i
llumination. Specifically, using the bidirectional flows estimated using an exis
ting pre-trained flow network, we predict the flows from an intermediate frame t
o the two input images. To do this, we propose to encode the bidirectional flows
 into a coordinate-based network, powered by a hypernetwork, to obtain a continu
ous representation of the flow across time. Once we obtain the estimated flows,
we use them within an existing blending network to obtain the final intermediate
 frame. Through extensive experiments, we demonstrate that our approach is able
to produce significantly better results than state-of-the-art frame interpolatio
n algorithms.
**********

Learning Lightweight Neural Networks via Channel-Split Recurrent Convolution
Guojun Wu, Xin Zhang, Ziming Zhang, Yanhua Li, Xun Zhou, Christopher Brinton, Zh
enming Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Com
puter Vision (WACV), 2023, pp. 3858-3868
Lightweight neural networks refer to deep networks with small numbers of paramet
ers, which are allowed to be implemented in resource-limited hardware such as em
bedded systems. To learn such lightweight networks effectively and efficiently,
in this paper we propose a novel convolutional layer, namely Channel-Split Recur
rent Convolution (CSR-Conv), where we split the output channels to generate data
 sequences with length T as the input to the recurrent layers with shared weight
s. As a consequence, we can construct lightweight convolutional networks by simp
ly replacing (some) linear convolutional layers with CSR-Conv layers. We prove t
hat under mild conditions the model size decreases with the rate of O(1 / T^2).
Empirically we demonstrate the state-of-the-art performance using VGG-16, ResNet
-50, ResNet-56, ResNet-110, DenseNet-40, MobileNet, and EfficientNet as backbone
 networks on CIFAR-10 and ImageNet. Codes can be found on https://github.com/tua
xon/CSR_Conv.
**********

Hear the Flow: Optical Flow-Based Self-Supervised Visual Sound Source Localizati
on
Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, Venu Gov
indaraju; Proceedings of the IEEE/CVF Winter Conference on Applications of Compu
ter Vision (WACV), 2023, pp. 2278-2287
Learning to localize the sound source in videos without explicit annotations is
a novel area of audio-visual research. Existing work in this area focuses on cre
ating attention maps to capture the correlation between the two modalities to lo
calize the source of the sound. In a video, oftentimes, the objects exhibiting m
ovement are the ones generating the sound. In this work, we capture this charact
eristic by modeling the optical flow in a video as a prior to better aid in loca
lizing the sound source. We further demonstrate that the addition of flow-based
attention substantially improves visual sound source localization. Finally, we b
enchmark our method on standard sound source localization datasets and achieve s
tate-of-the-art performance on the Soundnet Flickr and VGG Sound Source datasets
. Code: https://github.com/denfed/heartheflow.
**********

3DMM-RF: Convolutional Radiance Fields for 3D Face Modeling

Stathis Galanakis, Baris Gecer, Alexandros Lattas, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3536-3547

Facial 3D Morphable Models are a main computer vision subject with countless applications and have been highly optimized in the last two decades. The tremendous improvements of deep generative networks have created various possibilities for improving such models and have attracted wide interest. Moreover, the recent advances in neural radiance fields, are revolutionising novel-view synthesis of known scenes. In this work, we present a facial 3D Morphable Model, which exploits both of the above, and can accurately model a subject's identity, pose and expression and render it in arbitrary illumination. This is achieved by utilizing a powerful deep style-based generator to overcome two main weaknesses of neural radiance fields, their rigidity and rendering speed. We introduce a style-based generative network that synthesizes in one pass all and only the required rendering samples of a neural radiance field. We create a vast labelled synthetic dataset of facial renders, and train the network, so that it can accurately model and generalize on facial identity, pose and appearance. Finally, we show that this model can accurately be fit to "in-the-wild" facial images of arbitrary pose and illumination, extract the facial characteristics, and be used to re-render the face in controllable conditions.

**********

Avoiding Lingering in Learning Active Recognition by Adversarial Disturbance

Lei Fan, Ying Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4612-4621

This paper considers the active recognition scenario, where the agent is empowered to intelligently acquire observations for better recognition. The agents usually compose two modules, i.e., the policy and the recognizer, to select actions and predict the category. While using ground-truth class labels to supervise the recognizer, the policy is typically updated with rewards determined by the current in-training recognizer, like whether achieving correct predictions. However, this joint learning process could lead to unintended solutions, like a collapsed policy that only visits views that the recognizer is already sufficiently trained to obtain rewards, which harms the generalization ability. We call this phenomenon lingering to depict the agent being reluctant to explore challenging views during training. Existing approaches to tackle the exploration-exploitation trade-off could be ineffective as they usually assume reliable feedback during exploration to update the estimate of rarely-visited states. This assumption is invalid here as the reward from the recognizer could be insufficiently trained. To this end, our approach integrates another adversarial policy to constantly disturb the recognition agent during training, forming a competing game to promote active explorations and avoid lingering. The reinforced adversary, rewarded when the recognition fails, contests the recognition agent by turning the camera to challenging observations. Extensive experiments across two datasets validate the effectiveness of the proposed approach regarding its recognition performances, learning efficiencies, and especially robustness in managing environmental noises.

**********

Segmentation-Free Direct Iris Localization Networks

Takahiro Toizumi, Koichi Takahashi, Masato Tsukada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 991-1000

This paper proposes an efficient iris localization method without using iris segmentation and circle fitting. Conventional iris localization methods first extract iris regions by using semantic segmentation methods such as U-Net. Afterward, the inner and outer iris circles are localized using the traditional circle fitting algorithm. However, this approach requires high-resolution encoder-decoder networks for iris segmentation, so it causes computational costs to be high. In addition, traditional circle fitting tends to be sensitive to noise in input images and fitting parameters, causing the iris recognition performance to be poor. To solve these problems, we propose an iris localization network (ILN), that can directly localize pupil and iris circles with eyelid points from a low-resolution iris image. We also introduce a pupil refinement network (PRN) to improve th

e accuracy of pupil localization. Experimental results show that the combination of ILN and PRN works in 34.5 ms for one iris image on a CPU, and its localization performance outperforms conventional iris segmentation methods. In addition, generalized evaluation results show that the proposed method has higher robustness for datasets in different domain than other segmentation methods. Furthermore, we also confirm that the proposed ILN and PRN improve the iris recognition accuracy.

**********

Semantic Segmentation With Active Semi-Supervised Learning

Aneesh Rangnekar, Christopher Kanan, Matthew Hoffman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5966-5977

Using deep learning, we now have the ability to create exceptionally good semantic segmentation systems; however, collecting the prerequisite pixel-wise annotations for training images remains expensive and time-consuming. Therefore, it would be ideal to minimize the number of human annotations needed when creating a new dataset. Here, we address this problem by proposing a novel algorithm that combines active learning and semi-supervised learning. Active learning is an approach for identifying the best unlabeled samples to annotate. While there has been work on active learning for segmentation, most methods require annotating all pixel objects in each image, rather than only the most informative regions. We argue that this is inefficient. Instead, our active learning approach aims to minimize the number of annotations per image. Our method is enriched with semi-supervised learning, where we use pseudo labels generated with a teacher-student framework to identify image regions that help disambiguate confused classes. We also integrate mechanisms that enable better performance on imbalanced label distributions, which have not been studied previously for active learning in semantic segmentation. In experiments on the CamVid and CityScapes datasets, our method obtains over 95% of the network's performance on the full-training set using less than 17% of the training data, whereas the previous state of the art required 40% of the training data.

**********

Do Adaptive Active Attacks Pose Greater Risk Than Static Attacks?

Nathan Drenkow, Max Lennon, I-Jeng Wang, Philippe Burlina; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1380-1389

In contrast to perturbation-based attacks, patch-based attacks are physically realizable, and are therefore increasingly studied. However, prior work neglects the possibility of adaptive attacks optimized for 3D pose. For the first time, to our knowledge, we consider the challenge of designing and evaluating attacks on image sequences using 3D optimization along entire 3D kinematic trajectories. In this context, we study a type of dynamic attack, referred to as "adaptive active attacks" (AAA), that takes into consideration the pose of the observer being targeted. To better address the threat and risk posed by AAA attacks, we develop several novel risk-based and trajectory-based metrics. These are designed to capture the risk of attack success for attacking earlier in the trajectory to derail autonomous driving systems as well as tradeoffs that may arise given the possibility of additional detection. We evaluate performance of white-box targeted attacks using a subset of ImageNet classes, and demonstrate, in aggregate, that AAA attacks can pose threats beyond static attacks in kinematic settings in situations of predominantly looming motion (i.,e., a prevalent use case in automated vehicular navigation). Results demonstrate that AAA attacks can exhibit targeted attack success exceeding 10% in aggregate, and for some specific classes, up to 15% over their static counterparts. However, taking into consideration the probability of detection by the defender shows a more nuanced risk pattern. These new insights are important for guiding future adversarial machine learning studies and suggest researchers should consider defense against novel threats posed by dynamic attacks for full trajectories and videos.

**********

Improving Diversity With Adversarially Learned Transformations for Domain Genera

lization
Tejas Gokhale, Rushil Anirudh, Jayaraman J. Thiagarajan, Bhavya Kailkhura, Chitta Baral, Yezhou Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 434-443
To be successful in single source domain generalization (SSDG), maximizing diversity of synthesized domains has emerged as one of the most effective strategies. Recent success in SSDG comes from methods that pre-specify diversity inducing image augmentations during training, so that it may lead to better generalization on new domains. However, naive pre-specified augmentations are not always effective, either because they cannot model large domain shift, or because the specific choice of transforms may not cover the types of shifts commonly occurring in domain generalization. To address this issue, we present a novel framework called ALT: adversarially learned transformations, that uses an adversary neural network to model plausible, yet hard image transformations that fool the classifier. ALT learns image transformations by randomly initializing the adversary network for each batch and optimizing it for a fixed number of steps to maximize classification error. The classifier is trained by enforcing a consistency between its predictions on the clean and transformed images. With extensive empirical analysis, we find that this new form of adversarial transformations achieves both objectives of diversity and hardness simultaneously, outperforming all existing techniques on competitive benchmarks for SSDG. We also show that ALT can seamlessly work with existing diversity modules to produce highly distinct, and large transformations of the source domain leading to state-of-the-art performance. Code: https://github.com/tejas-gokhale/ALT
**********

Heatmap-Based Out-of-Distribution Detection
Julia Hornauer, Vasileios Belagiannis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2603-2612
Our work investigates out-of-distribution (OOD) detection as a neural network output explanation problem. We learn a heatmap representation for detecting OOD images while visualizing in- and out-of-distribution image regions at the same time. Given a trained and fixed classifier, we train a decoder neural network to produce heatmaps with zero response for in-distribution samples and high response heatmaps for OOD samples, based on the classifier features and the class prediction. Our main innovation lies in the heatmap definition for an OOD sample, as the normalized difference from the closest in-distribution sample. The heatmap serves as a margin to distinguish between in- and out-of-distribution samples. Our approach generates the heatmaps not only for OOD detection, but also to indicates in- and out-of-distribution regions of the input image. In our evaluations, our approach mostly outperforms the prior work on fixed classifiers, trained on CIFAR-10, CIFAR-100 and Tiny ImageNet. The code is publicly available at: https://github.com/jhornauer/heatmap_ood.
**********

Efficient Few-Shot Learning for Pixel-Precise Handwritten Document Layout Analysis
Axel De Nardin, Silvia Zottin, Matteo Paier, Gian Luca Foresti, Emanuela Colombi, Claudio Piciarelli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3680-3688
Layout analysis is a task of uttermost importance in ancient handwritten document analysis and represents a fundamental step toward the simplification of subsequent tasks such as optical character recognition and automatic transcription. However, many of the approaches adopted to solve this problem rely on a fully supervised learning paradigm. While these systems achieve very good performance on this task, the drawback is that pixel-precise text labeling of the entire training set is a very time-consuming process, which makes this type of information rarely available in a real-world scenario. In the present paper, we address this problem by proposing an efficient few-shot learning framework that achieves performances comparable to current state-of-the-art fully supervised methods on the publicly available DIVA-HisDB dataset
**********

Meta-Learning for Adaptation of Deep Optical Flow Networks
Chaerin Min, Taehyun Kim, Jongwoo Lim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2145-2154
In this paper, we propose an instance-wise meta-learning algorithm for optical flow domain adaptation. Typical optical flow algorithms with deep learning suffer from weak cross-domain performance since their trainings largely rely on synthetic datasets in specific domains. This prevents optical flow performance on different scenes from carrying similar performance in practice. Meanwhile, test-time domain adaptation approaches for optical flow estimation are yet to be studied. Our proposed method, with some training data, learns to adapt more sensitively to incoming inputs in the target domain. During the inference process, our method readily exploits the information only accessible in the test-time. Since our algorithm adapts to each input image, we incorporate traditional unsupervised losses for optical flow estimation. Moreover, with the observation that optical flows in a single domain typically contain many similar motions, we show that our method demonstrates high performance with only a small number of training data. This allows to save labeling efforts. Through the experiments on KITTI and MPI-Sintel datasets, our algorithm significantly outperforms the results without adaptation and shows consistently better performance in comparison to typical fine-tuning with the same amount of data. Also qualitatively our proposed method demonstrates more accurate results for the images with high errors in the original networks.
**********
Encouraging Disentangled and Convex Representation With Controllable Interpolation Regularization
Yunhao Ge, Zhi Xu, Yao Xiao, Gan Xin, Yunkui Pang, Laurent Itti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4761-4769
We focus on controllable disentangled representation learning (C-Dis-RL), where users can control the partition of the disentangled latent space to factorize dataset attributes (concepts) for downstream tasks. Two general problems remain under-explored in current methods: (1) They lack comprehensive disentanglement constraints, especially missing the minimization of mutual information between different attributes across latent and observation domains. (2) They lack convexity constraints, which is important for meaningfully manipulating specific attributes for downstream tasks. To encourage both comprehensive C-Dis-RL and convexity simultaneously, we propose a simple yet efficient method: Controllable Interpolation Regularization (CIR), which creates a positive loop where disentanglement and convexity can help each other. Specifically, we conduct controlled interpolation in latent space during training, and we reuse the encoder to help form a 'perfect disentanglement' regularization. In that case, (a) disentanglement loss implicitly enlarges the potential understandable distribution to encourage convexity; (b) convexity can in turn improve robust and precise disentanglement. CIR is a general module and we merge CIR with three different algorithms: ELEGANT, I2I-Dis, and GZS-Net to show the compatibility and effectiveness. Qualitative and quantitative experiments show improvement in C-Dis-RL and latent convexity by CIR. This further improves downstream tasks: controllable image synthesis, cross-modality image translation and zero-shot synthesis.
**********
Generative Alignment of Posterior Probabilities for Source-Free Domain Adaptation
Sachin Chhabra, Hemanth Venkateswara, Baoxin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4125-4134
Existing domain adaptation literature comprises multiple techniques that align the labeled source and unlabeled target domains at different stages, and predict the target labels. In a source-free domain adaptation setting, the source data is not available for alignment. We present a source-free generative paradigm that captures the relations between the source categories and enforces them onto the unlabeled target data, thereby circumventing the need for source data without introducing any new hyper-parameters. The adaptation is performed through the adv

ersarial alignment of the posterior probabilities of the source and target categ
ories. The proposed approach demonstrates competitive performance against other
source-free domain adaptation techniques and can also be used for source-present
 settings.
**********

DyStyle: Dynamic Neural Network for Multi-Attribute-Conditioned Style Editings
Bingchuan Li, Shaofei Cai, Wei Liu, Peng Zhang, Qian He, Miao Hua, Zili Yi; Proc
eedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WA
CV), 2023, pp. 189-197
The semantic controllability of StyleGAN is enhanced by unremitting research. Al
though the existing weak supervision methods work well in manipulating the style
 codes along one attribute, the accuracy of manipulating multiple attributes is
neglected. Multi-attribute representations are prone to entanglement in the Styl
eGAN latent space, while sequential editing leads to error accumulation. To addr
ess these limitations, we design a Dynamic Style Manipulation Network (DyStyle)
whose structure and parameters vary by input samples, to perform nonlinear and a
daptive manipulation of latent codes for flexible and precise attribute control.
 In order to efficient and stable optimization of the DyStyle network, we propos
e a Dynamic Multi-Attribute Contrastive Learning (DmaCL) method: including dynam
ic multi-attribute contrastor and dynamic multi-attribute contrastive loss, whic
h simultaneously disentangle a variety of attributes from the generative image a
nd latent space of model. As a result, our approach demonstrates fine-grained di
sentangled edits along multiple numeric and binary attributes. Qualitative and q
uantitative comparisons with existing style manipulation methods verify the supe
riority of our method in terms of the multi-attribute control accuracy and ident
ity preservation without compromising photorealism.
**********

Self-Attention Message Passing for Contrastive Few-Shot Learning
Ojas Kishorkumar Shirekar, Anuj Singh, Hadi Jamali-Rad; Proceedings of the IEEE/
CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5426-
5436
Humans have a unique ability to learn new representations from just a handful of
 examples with little to no supervision. Deep learning models, however, require
an abundance of data and supervision to perform at a satisfactory level. Unsuper
vised few-shot learning (U-FSL) is the pursuit of bridging this gap between mach
ines and humans. Inspired by the capacity of graph neural networks (GNNs) in dis
covering complex inter-sample relationships, we propose a novel self-attention b
ased message passing contrastive learning approach (coined as SAMP-CLR) for U-FS
L pre-training. We also propose an optimal transport (OT) based fine-tuning stra
tegy (we call OpT-Tune) to efficiently induce task awareness into our novel end-
to-end unsupervised few-shot classification framework (SAMPTransfer). Our extens
ive experimental results corroborate the efficacy of SAMPTransfer in a variety o
f downstream few-shot classification scenarios, setting a new state-of-the-art f
or U-FSL on both miniImageNet and tieredImageNet benchmarks, offering up to 7%+
and 5%+ improvements, respectively. Our further investigations also confirm that
 SAMPTransfer remains on-par with some supervised baselines on miniImageNet and
outperforms all existing U-FSL baselines in a challenging cross-domain scenario.
**********

Composite Relationship Fields With Transformers for Scene Graph Generation
George Adaimi, David Mizrahi, Alexandre Alahi; Proceedings of the IEEE/CVF Winte
r Conference on Applications of Computer Vision (WACV), 2023, pp. 52-64
Scene graph generation (SGG) methods extract relationships between objects. Whil
e most methods focus on improving top-down approaches, which build a scene graph
 based on detected objects from an off-the-shelf object detector, there is a lim
ited amount of work on bottom-up approaches, which jointly detect objects and th
eir relationships in a single stage. In this work, we present a novel bottom-up
SGG approach by representing relationships using Composite Relationship Fields (
CoRF). CoRF turns relationship detection into a dense regression and classificat
ion task, where each cell of the output feature map identifies surrounding objec
ts and their relationships. Furthermore, we propose a refinement head that lever

ages Transformers for global scene reasoning, resulting in more meaningful relat ionship predictions. By combining both contributions, our method outperforms pre vious bottom-up methods on the Visual Genome dataset by 26% while preserving rea l-time performance.
**********

ElliPose: Stereoscopic 3D Human Pose Estimation by Fitting Ellipsoids
Christian Grund, Julian Tanke, Jürgen Gall; Proceedings of the IEEE/CVF Winter C onference on Applications of Computer Vision (WACV), 2023, pp. 2871-2881
One of the most relevant tasks for augmented and virtual reality applications is the interaction of virtual objects with real humans which requires accurate 3D human pose predictions. Obtaining accurate 3D human poses requires careful camer a calibration which is difficult for non-technical personal or in a pop-up scena rio. Recent markerless motion capture approaches require accurate camera calibra tion at least for the final triangulation step. Instead, we solve this problem b y presenting ElliPose, Stereoscopic 3D Human Pose Estimation by Fitting Ellipsoi ds, where we jointly estimate the 3D human as well as the camera pose. We exploi t the fact that bones do not change in length over the course of a sequence and thus their relative trajectories have to lie on the surface of a sphere which we can utilize to iteratively correct the camera and 3D pose estimation. As anothe r use-case we demonstrate that our approach can be used as replacement for groun d-truth 3D poses to train monocular 3D pose estimators. We show that our method produces competitive results even when comparing with state-of-the-art methods t hat use more cameras or ground-truth camera extrinsics.
**********

Is Bigger Always Better? An Empirical Study on Efficient Architectures for Style Transfer and Beyond
Jie An, Tao Li, Haozhi Huang, Jinwen Ma, Jiebo Luo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4084-4094
Network architecture plays a pivotal role in the performance of style transfer a lgorithms. Most existing algorithms use VGG19 as the feature extractor, which in curs a high computational cost. In this work, we conduct an empirical study on t he popular network architectures and find that some more efficient networks can replace VGG19 while having comparable style transfer performance. Beyond that, w e show that an efficient network can be further accelerated by removing its empt y channels via a simple channel pruning method tweaked for style transfer. To pr event the potential performance drop due to using a more lightweight network and obtain better style transfer results, we introduce a more accurate deep feature alignment strategy to improve existing style transfer modules. Taking GoogLeNet as an exemplary efficient network, the pruned GoogLeNet with the improved style transfer module is 2.3 107.4x faster than the state-of-the-art approaches and can achieve 68.03 FPS on 512 x 512 images. Extensive experiments demonstrate th at VGG19 can be replaced by a more lightweight network with significantly improv ed efficiency and comparable style transfer quality.
**********

Watching the News: Towards VideoQA Models That Can Read
Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, C. V. Jawahar; Proceedin gs of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4441-4450
Video Question Answering methods focus on common-sense reasoning and visual cogn ition of objects or persons and their interactions over time. Current VideoQA ap proaches ignore the textual information present in the video. Instead, we argue that textual information is complementary to the action and provides essential c ontextualisation cues to the reasoning process. To this end, we propose a novel VideoQA task that requires reading and understanding the text in the video. To e xplore this direction, we focus on news videos and require QA systems to compreh end and answer questions about the topics presented by combining visual and text ual cues in the video. We introduce the "NewsVideoQA" dataset that comprises mor e than 8,600 QA pairs on 3, 000+ news videos obtained from diverse news channels from around the world. We demonstrate the limitations of current Scene Text VQA and VideoQA methods and propose ways to incorporate scene text information into

VideoQA methods.
**********
ATCON: Attention Consistency for Vision Models
Ali Mirzazadeh, Florian Dubost, Maxwell Pike, Krish Maniar, Max Zuo, Christopher Lee-Messer, Daniel Rubin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1880-1889
Attention--or attribution--maps methods are methods designed to highlight regions of the model's input that were discriminative for its predictions. However, different attention maps methods can highlight different regions of the input, with sometimes contradictory explanations for a prediction. This effect is exacerbated when the training set is small. This indicates that either the model learned incorrect representations or that the attention maps methods did not accurately estimate the model's representations. We propose an unsupervised fine-tuning method that optimizes the consistency of attention maps and show that it improves both classification performance and the quality of attention maps. We propose an implementation for two state-of-the-art attention computation methods, Grad-CAM and Guided Backpropagation, which relies on an input masking technique. We also show results on Grad-CAM and Integrated Gradients in an ablation study. We evaluate this method on our own dataset of event detection in continuous video recordings of hospital patients aggregated and curated for this work. As a sanity check, we also evaluate the proposed method on PASCAL VOC and SVHN. With the proposed method, with small training sets, we achieve a 6.6 points lift of F1 score over the baselines on our video dataset, a 2.9 point lift of F1 score on PASCAL, and a 1.8 points lift of mean Intersection over Union over Grad-CAM for weakly supervised detection on PASCAL. Those improved attention maps may help clinicians better understand vision model predictions and ease the deployment of machine learning systems into clinical care. We share part of the code for this article at the following repository: https://github.com/alimirzazadeh/SemisupervisedAttention.
**********
Computer Vision for Ocean Eddy Detection in Infrared Imagery
Evangelos Moschos, Alisa Kugusheva, Paul Coste, Alexandre Stegner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6395-6404
Reliable and precise detection of ocean eddies can significantly improve the monitoring of the ocean surface and subsurface dynamics, besides the characterization of local hydrographical and biological properties, or the concentration pelagic species. Today, most of the eddy detection algorithms operate on satellite altimetry gridded observations, which provide daily maps of sea surface height and surface geostrophic velocity. However, the reliability and the spatial resolution of altimetry products is limited by the strong spatio-temporal averaging of the mapping procedure. Yet, the availability of high-resolution satellite imagery makes real-time object detection possible at a much finer scale, via advanced computer vision methods. We propose a novel eddy detection method via a transfer learning schema, using the ground truth of high-resolution ocean numerical models to link the characteristic streamlines of eddies with their signature (gradients, swirls, and filaments) on Sea Surface Temperature (SST). A trained, multi-task convolutional neural network is then employed to segment infrared satellite imagery of SST in order to retreive the accurate position, size, and form of each detected eddy. The EddyScan-SST is an operational oceanographic module that provides, in real-time, key information on the ocean dynamics to maritime stakeholders.
**********
Expansion of Visual Hints for Improved Generalization in Stereo Matching
Andrea Pilzer, Yuxin Hou, Niki Loppi, Arno Solin, Juho Kannala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5840-5849
We introduce visual hints expansion for guiding stereo matching to improve generalization. Our work is motivated by the robustness of Visual Inertial Odometry (VIO) in computer vision and robotics, where a sparse and unevenly distributed se

t of feature points characterizes a scene. To improve stereo matching, we propos e to elevate 2D hints to 3D points. These sparse and unevenly distributed 3D vis ual hints are expanded using a 3D random geometric graph, which enhances the lea rning and inference process. We evaluate our proposal on multiple widely adopted benchmarks and show improved performance without access to additional sensors o ther than the image sequence. To highlight practical applicability and symbiosis with visual odometry, we demonstrate how our methods run on embedded hardware.
**********

Line Search-Based Feature Transformation for Fast, Stable, and Tunable Content-S tyle Control in Photorealistic Style Transfer
Tai-Yin Chiu, Danna Gurari; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2023, pp. 249-258
Photorealistic style transfer is the task of synthesizing a realistic-looking im age when adapting the content from one image to appear in the style of another i mage. Modern models commonly embed a transformation that fuses features describi ng the content image and style image and then decodes the resulting feature into a stylized image. We introduce a general-purpose transformation that enables co ntrolling the balance between how much content is preserved and the strength of the infused style. We offer the first experiments that demonstrate the performan ce of existing transformations across different style transfer models, and demon strate how transformation performs better in its ability to simultaneously run f ast, produce consistently reasonable results, and control the balance between co ntent and style in different models. To support reproducing our method and model s, we share the code at https://github.com/chiutaiyin/LS-FT.
**********

SketchInverter: Multi-Class Sketch-Based Image Generation via GAN Inversion
Zirui An, Jingbo Yu, Runtao Liu, Chuang Wang, Qian Yu; Proceedings of the IEEE/C VF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4319-4 329
This paper proposes the first GAN inversion-based method for multi-class sketch- based image generation (MC-SBIG). MC-SBIG is a challenging task that requires st rong prior knowledge due to the significant domain gap between sketches and natu ral images. Existing learning-based approaches rely on a large-scale paired data set to learn the mapping between these two image modalities. However, since the public paired sketch-photo data are scarce, it is struggling for learning-based methods to achieve satisfactory results. In this work, we introduce a new approa ch based on GAN inversion, which can utilize a powerful pretrained generator to facilitate image generation from a given sketch. Our GAN inversion-based method has two advantages: 1. it can freely take advantage of the prior knowledge of a pretrained image generator; 2. it allows the proposed model to focus on learning the mapping from a sketch to a low-dimension latent code, which is a much easie r task than directly mapping to a high-dimension natural image. We also present a novel shape loss to improve generation quality further. Extensive experiments are conducted to show that our method can produce sketch-faithful and photo-real istic images and significantly outperform the baseline methods.
**********

FAN-Trans: Online Knowledge Distillation for Facial Action Unit Detection
Jing Yang, Jie Shen, Yiming Lin, Yordan Hristov, Maja Pantic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6019-6027
Due to its importance in facial behaviour analysis, facial action unit (AU) dete ction has attracted increasing attention from the research community. Leveraging the online knowledge distillation framework, we propose the "FAN-Trans" method for AU detection. Our model consists of a hybrid network of convolution layers a nd transformer blocks designed to learn per-AU features and to model AU co-occur rences. The model uses a pre-trained face alignment network as the feature extra ctor. After further transformation by a small learnable add-on convolutional sub net, the per-AU features are fed into transformer blocks to enhance their repres entation. As multiple AUs often appear together, we propose a learnable attentio n drop mechanism in the transformer block to learn the correlation between the f

eatures for different AUs. We also design a classifier that predicts AU presence by considering all AUs' features, to explicitly capture label dependencies. Finally, we make the first attempt of adapting online knowledge distillation in the training stage for this task, further improving the model's performance. Experiments on the BP4D and DISFA datasets show our method has achieved a new state-of-the-art performance on both, demonstrating its effectiveness.
**********

Centroid Distance Keypoint Detector for Colored Point Clouds
Hanzhe Teng, Dimitrios Chatziparaschis, Xinyue Kan, Amit K. Roy-Chowdhury, Konstantinos Karydis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1196-1205
Keypoint detection serves as the basis for many computer vision and robotics applications. Despite the fact that colored point clouds can be readily obtained, most existing keypoint detectors extract only geometry-salient keypoints, which can impede the overall performance of systems that intend to (or have the potential to) leverage color information. To promote advances in such systems, we propose an efficient multi-modal keypoint detector that can extract both geometry-salient and color-salient keypoints in colored point clouds. The proposed CEntroid Distance (CED) keypoint detector comprises an intuitive and effective saliency measure, the centroid distance, that can be used in both 3D space and color space, and a multi-modal non-maximum suppression algorithm that can select keypoints with high saliency in two or more modalities. The proposed saliency measure leverages directly the distribution of points in a local neighborhood and does not require normal estimation or eigenvalue decomposition. We evaluate the proposed method in terms of repeatability and computational efficiency (i.e. running time) against state-of-the-art keypoint detectors on both synthetic and real-world datasets. Results demonstrate that our proposed CED keypoint detector requires minimal computational time while attaining high repeatability. To showcase one of the potential applications of the proposed method, we further investigate the task of colored point cloud registration. Results suggest that our proposed CED detector outperforms state-of-the-art handcrafted and learning-based keypoint detectors in the evaluated scenes. The C++ implementation of the proposed method is made publicly available at https://github.com/UCR-Robotics/CED_Detector.
**********

Seg&Struct: The Interplay Between Part Segmentation and Structure Inference for 3D Shape Parsing
Jeonghyun Kim, Kaichun Mo, Minhyuk Sung, Woontack Woo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1226-1235
We propose Seg&Struct, a supervised learning framework leveraging the interplay between part segmentation and structure inference and demonstrating their synergy in an integrated framework. Both part segmentation and structure inference have been extensively studied in the recent deep learning literature, while the supervisions used for each task have not been fully exploited to assist the other task. Namely, structure inference has been typically conducted with an autoencoder that does not leverage the point-to-part associations. Also, segmentation has been mostly performed without structural priors that tell the plausibility of the output segments. We present how these two tasks can be best combined while fully utilizing supervision to improve performance. Our framework first decomposes a raw input shape into part segments using an off-the-shelf algorithm, whose outputs are then mapped to nodes in a part hierarchy, establishing point-to-part associations. Following this, ours predicts the structural information, e.g., part bounding boxes and part relationships. Lastly, the segmentation is rectified by examining the confusion of part boundaries using the structure-based part features. Our experimental results based on the StructureNet and PartNet demonstrate that the interplay between the two tasks results in remarkable improvements in both tasks: 27.91% in structure inference and 0.5% in segmentation.
**********

Intention-Conditioned Long-Term Human Egocentric Action Anticipation
Esteve Valls Mascaró, Hyemin Ahn, Dongheui Lee; Proceedings of the IEEE/CVF Wint

er Conference on Applications of Computer Vision (WACV), 2023, pp. 6048-6057
To anticipate how a person would act in the future, it is essential to understand the human intention since it guides the subject towards a certain action. In this paper, we propose a hierarchical architecture which assumes a sequence of human action (low-level) can be driven from the human intention (high-level). Based on this, we deal with long-term action anticipation task in egocentric videos. Our framework first extracts this low- and high-level human information over the observed human actions in a video through a Hierarchical Multi-task Multi-Layer Perceptrons Mixer (H3M). Then, we constrain the uncertainty of the future through an Intention-Conditioned Variational Auto-Encoder (I-CVAE) that generates multiple stable predictions of the next actions that the observed human might perform. By leveraging human intention as high-level information, we claim that our model is able to anticipate more time-consistent actions in the long-term, thus improving the results over the baseline in Ego4D dataset. This work results in the state-of-the-art for Long-Term Anticipation (LTA) task in Ego4D by providing more plausible anticipated sequences, improving the anticipation scores of nouns and actions. Our work ranked first in both CVPR@2022 and ECCV@2022 Ego4D LTA Challenge.
**********
Learning Latent Structural Relations With Message Passing Prior
Shaogang Ren, Hongliang Fei, Dingcheng Li, Ping Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5334-5343
Learning disentangled representations is an important topic in machine learning with a wide range of applications. Disentangled latent variables represent interpretable semantic information and reflect separate factors of variation in data. Although generative models can learn latent representations as well, most existing models ignore the structural information among latent variables. In this paper, we propose a novel approach to learn the disentangled latent structural representations from data using decomposable variational auto-encoders. We design a novel message passing prior to the latent representations to capture the interactions among different data components. Different from many previous methods that ignore data component or object interaction, our approach simultaneously learns component representation and encodes component relationships. We have applied our model to tasks of data segmentation and latent representation learning among different data components. Experiments on several benchmarks demonstrate the utility of the proposed method.
**********
TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos
Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 137-146
Weakly supervised video object localization (WSVOL) allows locating object in videos using only global video tags such as object classes. State-of-art methods rely on multiple independent stages, where initial spatio-temporal proposals are generated using visual and motion cues, and then prominent objects are identified and refined. The localization involves solving an optimization problem over one or more videos, and video tags are typically used for video clustering. This process requires a model per video or per class making for costly inference. Moreover, localized regions are not necessary discriminant because these methods rely on unsupervised motion methods like optical flow, or discarded video tags from optimization. In this paper, we leverage the successful class activation mapping (CAM) methods, designed for WSOL based on still images. A new Temporal CAM (TCAM) method is introduced for training a discriminant deep learning (DL) model to exploit spatio-temporal information in videos, using an CAM-Temporal Max Pooling (CAM-TMP) aggregation mechanism over consecutive CAMs. In particular, activations of regions of interest (ROIs) are collected from CAMs produced by a pretrained CNN classifier, and generate pixel-wise pseudo-labels for training a decoder. In addition, a global unsupervised size constraint, and local constraint such as CRF are used to yield more accurate CAMs. Inference over single independent fr

ames allows parallel processing of a clip of frames, and real-time localization. Extensive experiments on two challenging YouTube-Objects datasets with unconstrained videos indicate that CAM methods (trained on independent frames) can yield decent localization accuracy. Our proposed TCAM method achieves a new state-of-art in WSVOL accuracy, and visual results suggest that it can be adapted for subsequent tasks, such as object detection and tracking.

**********

Bootstrapping the Relationship Between Images and Their Clean and Noisy Labels

Brandon Smart, Gustavo Carneiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5344-5354

Many state-of-the-art noisy-label learning methods rely on learning mechanisms that estimate the samples' clean labels during training and discard their original noisy labels. However, this approach prevents the learning of the relationship between images, noisy labels and clean labels, which has been shown to be useful when dealing with instance-dependent label noise problems. Furthermore, methods that do aim to learn this relationship require cleanly annotated subsets of data, as well as distillation or multi-faceted models for training. In this paper, we propose a new training algorithm that relies on a simple model to learn the relationship between clean and noisy labels without the need for a cleanly labelled subset of data. Our algorithm follows a 3-stage process, namely: 1) self-supervised pretraining followed by an early-stopping training of the classifier to confidently predict clean labels for a subset of the training set; 2) use the clean set from stage (1) to bootstrap the relationship between images, noisy labels and clean labels, which we exploit for effective relabelling of the remaining training set using semi-supervised learning; and 3) supervised training of the classifier with all relabelled samples from stage (2). By learning this relationship, we achieve state-of-the-art performance in asymmetric and instance-dependent label noise problems. Code is available at https://github.com/btsmart/bootstrapping-label-noise

**********

Full Contextual Attention for Multi-Resolution Transformers in Semantic Segmentation

Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, Alexandre Hostettler; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3224-3233

Transformers have proved to be very effective for visual recognition tasks. In particular, vision transformers construct compressed global representation through self-attention and learnable class tokens. Multi-resolution transformers have shown recent successes in semantic segmentation but can only capture local interactions in high-resolution feature maps. This paper extends the notion of global tokens to build GLobal Attention Multi-resolution (GLAM) transformers. GLAM is a generic module that can be integrated into most existing transformer backbones. GLAM includes learnable global tokens, which unlike previous methods can model interactions between all image regions, and extracts powerful representations during training. Extensive experiments show that GLAM-Swin or GLAM-Swin-Unet exhibit substantially better performances than their vanilla counterparts on ADE20K and Cityscapes. Moreover, GLAM can be used to segment large 3D medical images, and GLAM-nnFormer achieves new state-of-the-art performance on the BCV dataset.

**********

Adversarial Local Distribution Regularization for Knowledge Distillation

Thanh Nguyen-Duc, Trung Le, He Zhao, Jianfei Cai, Dinh Phung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4681-4690

Knowledge distillation is a process of distilling information from a large model with significant knowledge capacity (teacher) to enhance a smaller model (student). Therefore, exploring the properties of the teacher is the key to improving student performance (e.g., teacher decision boundaries). One decision boundary exploring technique is to leverage adversarial attack methods, which add crafted perturbations within a ball constraint to clean inputs to create attack examples of the teacher called adversarial examples. These adversarial examples are info

rmative examples because they are near decision boundaries. In this paper, we fo
rmulate a teacher adversarial local distribution, a set of all adversarial examp
les within the ball constraint given an input. This distribution is used to suff
iciently explore the decision boundaries of the teacher by covering the full spe
ctrum of possible teacher model perturbations. The student model is then regular
ized by matching the loss between teacher and student using these adversarial ex
ample inputs. We conducted a number of experiments on CIFAR-100 and Imagenet dat
asets to illustrate this teacher adversarial local distribution regularization (
TALD) can be applied to improve performance of many existing knowledge distillat
ion methods (e.g., KD, FitNet, CRD, VID, FT, etc.).
**********

Splatting-Based Synthesis for Video Frame Interpolation
Simon Niklaus, Ping Hu, Jiawen Chen; Proceedings of the IEEE/CVF Winter Conferen
ce on Applications of Computer Vision (WACV), 2023, pp. 713-723
Frame interpolation is an essential video processing technique that adjusts the
temporal resolution of an image sequence. While deep learning has brought great
improvements to the area of video frame interpolation, techniques that make use
of neural networks can typically not easily be deployed in practical application
s like a video editor since they are either computationally too demanding or fai
l at high resolutions. In contrast, we propose a deep learning approach that sol
ely relies on splatting to synthesize interpolated frames. This splatting-based
synthesis for video frame interpolation is not only much faster than similar app
roaches, especially for multi-frame interpolation, but can also yield new state-
of-the-art results at high resolutions.
**********

CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning-Based Synth
etic Face Detection
Aidan Boyd, Patrick Tinsley, Kevin W. Bowyer, Adam Czajka; Proceedings of the IE
EE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 61
08-6117
Can deep learning models achieve greater generalization if their training is gui
ded by reference to human perceptual abilities? And how can we implement this in
 a practical manner? This paper proposes a training strategy to ConveY Brain Ove
rsight to Raise Generalization (CYBORG). This new approach incorporates human-an
notated saliency maps into a loss function that guides the model's learning to f
ocus on image regions that humans deem salient for the task. The Class Activatio
n Mapping (CAM) mechanism is used to probe the model's current saliency in each
training batch, juxtapose this model saliency with human saliency, and penalize
large differences. Results on the task of synthetic face detection, selected to
illustrate the effectiveness of the approach, show that CYBORG leads to signific
ant improvement in accuracy on unseen samples consisting of face images generate
d from six Generative Adversarial Networks across multiple classification networ
k architectures. We also show that scaling to even seven times the training data
, or using non-human-saliency auxiliary information, such as segmentation masks,
 and standard loss cannot beat the performance of CYBORG-trained models. As a si
de effect of this work, we observe that the addition of explicit region annotati
on to the task of synthetic face detection increased human classification accura
cy. This work opens a new area of research on how to incorporate human visual sa
liency into loss functions in practice. All data, code and pre-trained models us
ed in this work are offered with this paper.
**********

On Quantizing Implicit Neural Representations
Cameron Gordon, Shin-Fang Chng, Lachlan MacDonald, Simon Lucey; Proceedings of t
he IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, p
p. 341-350
The role of quantization within implicit/coordinate neural networks is still not
 fully understood. We note that using a canonical fixed quantization scheme duri
ng training produces poor performance at low bit-rates due to the network weight
 distributions changing over the course of training. In this work, we show that
a non-uniform quantization of neural weights can lead to significant improvement

s. Specifically, we demonstrate that a clustered quantization enables improved r
econstruction. Finally, by characterising a trade-off between quantization and n
etwork capacity, we demonstrate that it is possible (while memory inefficient) t
o reconstruct signals using binary neural networks. We demonstrate our findings
experimentally on 2D image reconstruction and 3D radiance fields; and show that
simple quantization methods and architecture search can achieve compression of N
eRF to less than 16kb with minimal loss in performance (323x smaller than the or
iginal NeRF).
**********

ImPosing: Implicit Pose Encoding for Efficient Visual Localization
Arthur Moreau, Thomas Gilles, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciules
cu, Arnaud de La Fortelle; Proceedings of the IEEE/CVF Winter Conference on Appl
ications of Computer Vision (WACV), 2023, pp. 2892-2902
We propose a novel learning-based formulation for visual localization of vehicle
s that can operate in real-time in city-scale environments. Visual localization
algorithms determine the position and orientation from which an image has been c
aptured, using a set of geo-referenced images or a 3D scene representation. Our
new localization paradigm, named Implicit Pose Encoding (ImPosing), embeds image
s and camera poses into a common latent representation with 2 separate neural ne
tworks, such that we can compute a similarity score for each image-pose pair. By
 evaluating candidates through the latent space in a hierarchical manner, the ca
mera position and orientation are not directly regressed but incrementally refin
ed. Very large environments force competitors to store gigabytes of map data, wh
ereas our method is very compact independently of the reference database size. I
n this paper, we describe how to effectively optimize our learned modules, how t
o combine them to achieve real-time localization, and demonstrate results on div
erse large scale scenarios that significantly outperform prior work in accuracy
and computational efficiency.
**********

A Protocol for Evaluating Model Interpretation Methods From Visual Explanations
Hamed Behzadi-Khormouji, José Oramas; Proceedings of the IEEE/CVF Winter Confere
nce on Applications of Computer Vision (WACV), 2023, pp. 1421-1429
With the continuous development of Convolutional Neural Networks (CNNs), there i
s an increasing requirement towards the understanding of the representations the
y internally encode. The task of studying such encoded representations is referr
ed to as model interpretation. Efforts along this direction, despite being prove
d efficient, stand with two weaknesses. First, there is low semanticity on the f
eedback they provide which leads toward subjective visualizations. Second, there
 is no unified protocol for the quantitative evaluation of interpretation method
s which makes the comparison between current and future methods complex.\nTo add
ress these issues, we propose a unified evaluation protocol for the quantitative
 evaluation of interpretation methods. This is achieved by enhancing existing in
terpretation methods to be capable of generating visual explanations and then li
nking these explanations with a semantic label. To achieve this, we introduce th
e Weighted Average Intersection-over-Union (WAIoU) metric to estimate the covera
ge rate between explanation heatmaps and semantic annotations. This is complemen
ted with an analysis of several binarization techniques for heatmaps, necessary
when measuring coverage. Experiments considering several interpretation methods
covering different CNN architectures pre-trained on multiple datasets show the e
ffectiveness of the proposed protocol.
**********

Wiener Guided DIP for Unsupervised Blind Image Deconvolution
Gustav Bredell, Ertunc Erdil, Bruno Weber, Ender Konukoglu; Proceedings of the I
EEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3
047-3056
Blind deconvolution is an ill-posed problem arising in various fields ranging fr
om microscopy to astronomy. Its ill-posed nature demands adequate priors and ini
tialization to arrive at a desirable solution. Recently, it has been shown that
deep networks can serve as an image generation prior (DIP) during unsupervised b
lind deconvolution optimization, however, DIP's high frequency artifact suppress

ion ability is not explicitly exploited. We propose to use Wiener-deconvolution to guide DIP during optimization in order to better leverage DIP's ability for blind image deconvolution. Wiener-deconvolution sharpens an image while introducing high-frequency artifacts, which are reproduced by DIP with a delay compared to low-frequency features and sharp edges, similar to what has been observed for noise. We embed the computational process in a constrained optimization problem together with an automatic kernel initialization method and show that the proposed method yields higher performance and stability across multiple datasets.
**********

GEMS: Scene Expansion Using Generative Models of Graphs
Rishi Agarwal, Tirupati Saketh Chandra, Vaidehi Patil, Aniruddha Mahapatra, Kuldeep Kulkarni, Vishwa Vinay; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 157-166
Applications based on image retrieval require editing and associating in intermediate spaces that are representative of the high-level concepts like objects and their relationships rather than dense, pixel-level representations like RGB images or semantic-label maps. We focus on one such representation, scene graphs, and propose a novel scene expansion task where we enrich an input seed graph by adding new nodes (objects) and the corresponding relationships. To this end, we formulate scene graph expansion as a sequential prediction task involving multiple iterations of first predicting a new node and then predicting the set of relationships between the newly predicted node and previously chosen nodes in the graph. We propose and evaluate a sequencing strategy that retains the clustering patterns amongst nodes. In addition, we leverage external knowledge to train our graph generation model, enabling greater generalization of node predictions. Due to the inefficiency of existing maximum mean discrepancy (MMD) based metrics standard for graph generation problems, we design novel metrics that comprehensively evaluate different aspects of node and relation predictions. We conduct extensive experiments on Visual Genome and VRD datasets to evaluate the expanded scene graphs using the standard MMD based metrics and our proposed metrics. We observe that the graphs generated by our method, GEMS, better represent the real distribution of the scene graphs compared with baseline methods like GraphRNN.
**********

Uncertainty-Aware Label Distribution Learning for Facial Expression Recognition
Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, Anh Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6088-6097
Despite significant progress over the past few years, ambiguity is still a key challenge in Facial Expression Recognition (FER). It can lead to noisy and inconsistent annotation, which hinders the performance of deep learning models in real-world scenarios. In this paper, we propose a new uncertainty-aware label distribution learning method to improve the robustness of deep models against uncertainty and ambiguity. We leverage neighborhood information in the valence-arousal space to adaptively construct emotiona distributions for training samples. We also consider the uncertainty of provided labels when incorporating them into the label distributions. Our method can be easily integrated into a deep network to obtain more training supervision and improve recognition accuracy. Intensive experiments on several datasets under various noisy and ambiguous settings show that our method achieves competitive results and outperforms recent state-of-the-art approaches. Our code and models are available at https://github.com/minhnhatvt/label-distribution-learning-fer-tf.
**********

Self-Supervised Monocular Depth Estimation From Thermal Images via Adversarial Multi-Spectral Adaptation
Ukcheol Shin, Kwanyong Park, Byeong-Uk Lee, Kyunghyun Lee, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5798-5807
Recently, thermal image based 3D understanding is gradually attracting attention for an illumination condition agnostic machine vision. However, the difficulty of the thermal image lies in insufficient training supervision due to its low-co

ntrast and textureless properties. Also, introducing additional modality require s further constraints such as complicated multi-sensor calibration and synchroni zed data acquisition. To leverage additional modality information without such c onstraints, we propose a novel training framework that consists of self-supervis ed learning of unpaired multi-spectral images and feature-level adversarial adap tation. In the training stage, we utilize unpaired RGB/thermal video and partial ly shared network architecture consisting of modality-specific feature extractor s and modality-independent decoder. Through the shared network design, the depth decoder can leverage the self-supervised signal of the unpaired RGB images. Fea ture-level adversarial adaptation minimizes the gap between RGB and thermal feat ures and eventually makes the thermal encoder extract representative and informa tive features. Based on the proposed method, the trained depth network shows out performed results than previous state-of-the-art methods.
**********

Single-Image HDR Reconstruction by Multi-Exposure Generation
Phuoc-Hieu Le, Quynh Le, Rang Nguyen, Binh-Son Hua; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4063-4072
High dynamic range (HDR) imaging is an indispensable technique in modern photogr aphy. Traditional methods focus on HDR reconstruction from multiple images, solv ing the core problems of image alignment, fusion, and tone mapping, yet having a perfect solution due to ghosting and other visual artifacts in the reconstructi on. Recent attempts at single-image HDR reconstruction show a promising alternat ive: by learning to map pixel values to their irradiance using a neural network, one can bypass the align-and-merge pipeline completely yet still obtain a high- quality HDR image. In this work, we propose a weakly supervised learning method that inverts the physical image formation process for HDR reconstruction via lea rning to generate multiple exposures from a single image. Our neural network can invert the camera response to reconstruct pixel irradiance before synthesizing multiple exposures and hallucinating details in under- and over-exposed regions from a single input image. To train the network, we propose a representation los s, a reconstruction loss, and a perceptual loss applied on pairs of under- and o ver-exposure images and thus do not require HDR images for training. Our experim ents show that our proposed model can effectively reconstruct HDR images. Our qu alitative and quantitative results show that our method achieves state-of-the-ar t performance on the DrTMO dataset. Our code is available at https://github.com/ VinAIResearch/single_image_hdr.
**********

Scaling Novel Object Detection With Weakly Supervised Detection Transformers
Tyler LaBonte, Yale Song, Xin Wang, Vibhav Vineet, Neel Joshi; Proceedings of th e IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp . 85-96
A critical object detection task is finetuning an existing model to detect novel objects, but the standard workflow requires bounding box annotations which are time-consuming and expensive to collect. Weakly supervised object detection (WSO D) offers an appealing alternative, where object detectors can be trained using image-level labels. However, the practical application of current WSOD models is limited, as they only operate at small data scales and require multiple rounds of training and refinement. To address this, we propose the Weakly Supervised De tection Transformer, which enables efficient knowledge transfer from a large-sca le pretraining dataset to WSOD finetuning on hundreds of novel objects. Addition ally, we leverage pretrained knowledge to improve the multiple instance learning (MIL) framework often used in WSOD methods. Our experiments show that our appro ach outperforms previous state-of-the-art models on large-scale novel object det ection datasets, and our scaling study reveals that class quantity is more impor tant than image quantity for WSOD pretraining.
**********

Auxiliary Task-Guided CycleGAN for Black-Box Model Domain Adaptation
Michael Essich, Markus Rehmann, Cristóbal Curio; Proceedings of the IEEE/CVF Win ter Conference on Applications of Computer Vision (WACV), 2023, pp. 541-550
The research area of domain adaptation investigates methods that enable the tran

sfer of existing models across different domains, e.g., addressing environmental changes or the transfer from synthetic to real data. Especially unsupervised domain adaptation is beneficial because it does not require any labeled target domain data. Usually, existing methods are targeted at specific tasks and require access or even modifications to the source model and its parameters which is a major drawback when only a black-box model is available. Therefore, we propose a CycleGAN-based approach suitable for black-box source models to translate target domain data into the source domain on which the source model can operate. Inspired by multi-task learning, we extend CycleGAN with an additional auxiliary task that can be arbitrarily chosen to support the transfer of task-related information across domains without the need for having access to a differentiable source model or its parameters. In this work, we focus on the regression task of 2D human pose estimation and compare our results in four different domain adaptation settings to CycleGAN and RegDA, a state-of-the-art method for unsupervised domain adaptation for keypoint detection.

**********

DeformIrisNet: An Identity-Preserving Model of Iris Texture Deformation

Siamul Karim Khan, Patrick Tinsley, Adam Czajka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 900-908

Nonlinear iris texture deformations due to pupil size variations are one of the main factors responsible for within-class variance of genuine comparison scores in iris recognition. In dominant approaches to iris recognition, the size of a ring-shaped iris region is linearly scaled to a canonical rectangle, used further in encoding and matching. However, the biological complexity of the iris sphincter and dilator muscles causes the movements of iris features to be nonlinear in a function of pupil size, and not solely organized along radial paths. Alternatively to the existing theoretical models based on the biomechanics of iris musculature, in this paper we propose a novel deep autoencoder-based model that can effectively learn complex movements of iris texture features directly from the data. The proposed model takes two inputs, (a) an ISO-compliant near-infrared iris image with initial pupil size, and (b) the binary mask defining the target shape of the iris. The model makes all the necessary nonlinear deformations to the iris texture to match the shape of the iris in an image (a) with the shape provided by the target mask (b). The identity-preservation component of the loss function helps the model in finding deformations that preserve identity and not only the visual realism of the generated samples. We also demonstrate two immediate applications of this model: better compensation for iris texture deformations in iris recognition algorithms, compared to linear models, and the creation of a generative algorithm that can aid human forensic examiners, who may need to compare iris images with a large difference in pupil dilation. We offer the source codes and model weights available along with this paper.

**********

Task Agnostic and Post-Hoc Unseen Distribution Detection

Radhika Dua, Seongjun Yang, Yixuan Li, Edward Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1350-1359

Despite the recent advances in out-of-distribution(OOD) detection, anomaly detection, and uncertainty estimation tasks, there do not exist a task-agnostic and post-hoc approach. To address this limitation, we design a novel clustering-based ensembling method, called Task Agnostic and Post-hoc Unseen Distribution Detection (TAPUDD) that utilizes the features extracted from the model trained on a specific task. Explicitly, it comprises of TAP-Mahalanobis, which clusters the training datasets' features and determines the minimum Mahalanobis distance of the test sample from all clusters. Further, we propose the Ensembling module that aggregates the computation of iterative TAP-Mahalanobis for a different number of clusters to provide reliable and efficient cluster computation. Through extensive experiments on synthetic and real-world datasets, we observe that our task-agnostic approach can detect unseen samples effectively across diverse tasks and performs better or on-par with the existing task-specific baselines. We also demonstrate that our method is more viable even for large-scale classification tasks.

**********

SeCo: Separating Unknown Musical Visual Sounds With Consistency Guidance
Xinchi Zhou, Dongzhan Zhou, Wanli Ouyang, Hang Zhou, Di Hu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5168-5177
Recent years have witnessed the success of deep learning on the visual sound separation task. However, existing works follow similar settings where the training and testing datasets share the same musical instrument categories, which to some extent limits the versatility of this task. In this work, we focus on a more general and challenging scenario, namely the separation of unknown musical instruments, where the categories in training and testing phases have no overlap with each other. To tackle this new setting, we propose the "Separation-with-Consistency" (SeCo) framework, which can accomplish the separation on unknown categories by exploiting the consistency constraints. Furthermore, to capture richer characteristics of the novel melodies, we devise an online matching strategy, which can bring stable enhancements with no cost of extra parameters. Experiments demonstrate that our SeCo framework exhibits strong adaptation ability on the novel musical categories and outperforms the baseline methods by a notable margin.
**********

More Than Just Attention: Improving Cross-Modal Attentions With Contrastive Constraints for Image-Text Matching
Yuxiao Chen, Jianbo Yuan, Long Zhao, Tianlang Chen, Rui Luo, Larry Davis, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4432-4440
Cross-modal attention mechanisms have been widely applied to the image-text matching task and have achieved remarkable improvements thanks to their capability of learning fine-grained relevance across different modalities. However, the cross-modal attention models of existing methods could be sub-optimal and inaccurate because there is no direct supervision provided during the training process. In this work, we propose two novel training strategies, namely Contrastive Content Re-sourcing (CCR) and Contrastive Content Swapping (CCS) constraints, to address such limitations. These constraints supervise the training of cross-modal attention models in a contrastive learning manner without requiring explicit attention annotations. They are plug-in training strategies and can be generally integrated into existing cross-modal attention models. Additionally, we introduce three metrics, including Attention Precision, Recall, and F1-Score, to quantitatively measure the quality of learned attention models. We evaluate the proposed constraints by incorporating them into four state-of-the-art cross-modal attention-based image-text matching models. Experimental results on both Flickr30k and MS-COCO datasets demonstrate that integrating these constraints generally improves the model performance in terms of both retrieval performance and attention metrics.
**********

Autoencoder-Based Background Reconstruction and Foreground Segmentation With Background Noise Estimation
Bruno Sauvalle, Arnaud de La Fortelle; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3244-3255
Even after decades of research, dynamic scene background reconstruction and foreground object segmentation are still considered as open problems due to various challenges such as illumination changes, camera movements, or background noise caused by air turbulence or moving trees. We propose in this paper to model the background of a frame sequence as a low dimensional manifold using an autoencoder and compare the reconstructed background provided by this autoencoder with the original image to compute the foreground/background segmentation masks. The main novelty of the proposed model is that the autoencoder is also trained to predict the background noise, which allows to compute for each frame a pixel-dependent threshold to perform the foreground segmentation. Although the proposed model does not use any temporal or motion information, it exceeds the state of the art for unsupervised background subtraction on the CDnet 2014 and LASIESTA datasets, with a significant improvement on videos where the camera is moving. It is also able to perform background reconstruction on some non-video image datasets.

**********
Learning Few-Shot Segmentation From Bounding Box Annotations
Byeolyi Han, Tae-Hyun Oh; Proceedings of the IEEE/CVF Winter Conference on Appli cations of Computer Vision (WACV), 2023, pp. 3750-3759
We present a new weakly-supervised few-shot semantic segmentation setting and a meta-learning method for tackling the new challenge. Different from existing set tings, we leverage bounding box annotations as weak supervision signals during t he meta-training phase, i.e., more label-efficient. Bounding box provides a chea per label representation than segmentation mask but contains both an object of i nterest and a disturbing background. We first show that meta-training with bound ing boxes degrades recent few-shot semantic segmentation methods, which are typi cally meta-trained with full semantic segmentation supervision. We postulate tha t this challenge is originated from the impure information of bounding box repre sentation. We propose a pseudo trimap estimator and trimap-attention based proto type learning to extract clearer supervision signals from bounding boxes. These developments robustify and generalize our method well to noisy support masks at test time. We empirically show that our method consistently improves performance . Our method gains 1.4% and 3.6% mean-IoU over the competing one in full and wea k test supervision cases, respectively, in the 1-way 5-shot setting on Pascal-5i .
**********
AudioViewer: Learning To Visualize Sounds
Chunjin Song, Yuchi Zhang, Willis Peng, Parmis Mohaghegh, Bastian Wandt, Helge R hodin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2206-2216
A long-standing goal in the field of sensory substitution is enabling sound perc eption for deaf and hard of hearing (DHH) people by visualizing audio content. D ifferent from existing models that translate to hand sign language, between spee ch and text, or text and images, we target immediate and low-level audio to vide o translation that applies to generic environment sounds as well as human speech . Since such a substitution is artificial, without labels for supervised learnin g, our core contribution is to build a mapping from audio to video that learns f rom unpaired examples via high-level constraints. For speech, we additionally di sentangle content from style, such as gender and dialect. Qualitative and quanti tative results, including a human study, demonstrate that our unpaired translati on approach maintains important audio features in the generated video and that v ideos of faces and numbers are well suited for visualizing high-dimensional audi o features that can be parsed by humans to match and distinguish between sounds and words. Project website: https://chunjinsong.github.io/audioviewer
**********
Motion Aware Self-Supervision for Generic Event Boundary Detection
Ayush K. Rai, Tarun Krishna, Julia Dietlmeier, Kevin McGuinness, Alan F. Smeaton , Noel E. O'Connor; Proceedings of the IEEE/CVF Winter Conference on Application s of Computer Vision (WACV), 2023, pp. 2728-2739
The task of Generic Event Boundary Detection (GEBD) aims to detect moments in vi deos that are naturally perceived by humans as generic and taxonomy-free event b oundaries. Modeling the dynamically evolving temporal and spatial changes in a v ideo makes GEBD a difficult problem to solve. Existing approaches involve very c omplex and sophisticated pipelines in terms of architectural design choices, hen ce creating a need for more straightforward and simplified approaches. In this w ork, we address this issue by revisiting a simple and effective self-supervised method and augment it with a differentiable motion feature learning module to ta ckle the spatial and temporal diversities in the GEBD task. We perform extensive experiments on the challenging Kinetics-GEBD and TAPOS datasets to demonstrate the efficacy of the proposed approach compared to the other self-supervised stat e-of-the-art methods. We also show that this simple self-supervised approach lea rns motion features without any explicit motion-specific pretext task.
**********
Modeling Stroke Mask for End-to-End Text Erasing
Xiangcheng Du, Zhao Zhou, Yingbin Zheng, Tianlong Ma, Xingjiao Wu, Cheng Jin; Pr

oceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision ( WACV), 2023, pp. 6151-6159
Scene text erasing aims to wipe text regions in scene images with reasonable bac kground. Most previous approaches employ scene text detectors to assist localiza tion of the text regions. However, detected text boxes contain both text strokes and background clutters, and directly inpainting on the whole boxes may remain text artifacts and make regions unnatural. In this paper, we present an end-to-e nd network that focuses on modeling text stroke masks that provide more accurate locations to compute erased images. The network consists of two stages, i.e., a basic network with stroke generation and a refinement network with stroke aware ness. The basic network predicts the text stroke masks and initial erasing resul ts simultaneously. The refinement network receives the masks as supervision to g enerate natural erased results. Experiments on both synthetic and real-world sce ne images demonstrate the effectiveness of our framework in producing high quali ty erasing results.
**********

Mutual Learning for Long-Tailed Recognition
Changhwa Park, Junho Yim, Eunji Jun; Proceedings of the IEEE/CVF Winter Conferen ce on Applications of Computer Vision (WACV), 2023, pp. 2675-2684
Deep neural networks perform well in artificially-balanced datasets, but real-wo rld data often has a long-tailed distribution. Recent studies have focused on de veloping unbiased classifiers to improve tail class performance. Despite the eff orts to learn a fine classifier, we cannot guarantee a solid performance if the representations are of poor quality. However, learning high-quality representati ons in a long-tailed setting is difficult because the features of tail classes e asily overfit the training dataset. In this work, we propose a mutual learning f ramework that generates high-quality representations in long-tailed settings by exchanging information between networks. We show that the proposed method can im prove representation quality and establish a new state-of-the-art record on seve ral long-tailed recognition benchmark datasets, including CIFAR100-LT, ImageNet- LT, and iNaturalist 2018.
**********

GLAD: A Global-to-Local Anomaly Detector
Aitor Artola, Yannis Kolodziej, Jean-Michel Morel, Thibaud Ehret; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5501-5510
Learning to detect automatic anomalies in production plants remains a machine le arning challenge. Since anomalies by definition cannot be learned, their detecti on must rely on a very accurate "normality model". To this aim, we introduce her e a global-to-local Gaussian model for neural network features, learned from a s et of normal images. This probabilistic model enables unsupervised anomaly detec tion. A global Gaussian mixture model of the features is first learned using all available features from normal data. This global Gaussian mixture model is then localized by an adaptation of the K-MLE algorithm, which learns a spatial weigh t map for each Gaussian. These weights are then used instead of the mixture weig hts to detect anomalies. This method enables precise modeling of complex data, e ven with limited data. Applied on WideResnet50-2 features, our approach outperfo rms the previous state of the art on the MVTec dataset, particularly on the obje ct category. It is robust to perturbations that are frequent in production lines , such as imperfect alignment, and is on par in terms of memory and computation time with the previous state of the art.
**********

SPIQ: Data-Free Per-Channel Static Input Quantization
Edouard Yvinec, Arnaud Dapogny, Matthieu Cord, Kevin Bailly; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3869-3878
Computationally expensive neural networks are ubiquitous in computer vision and solutions for efficient inference have drawn a growing attention in the machine learning community. Examples of such solutions comprise quantization, i.e. conve rting the processing values (weights and inputs) from floating point into intege

rs e.g. int8 or int4. Concurrently, the rise of privacy concerns motivated the s tudy of less invasive acceleration methods, such as data-free quantization of pr e-trained models weights and activations. Previous approaches either exploit sta tistical information to deduce scalar ranges and scaling factors for the activat ions in a static manner, or dynamically adapt this range on-the-fly for each inp ut of each layers (also referred to as activations): the latter generally being more accurate at the expanse of significantly slower inference. In this work, we argue that static input quantization can reach the accuracy levels of dynamic m ethods by means of a per-channel input quantization scheme that allows one to mo re finely preserve cross-channel dynamics. We show through a thorough empirical evaluation on multiple computer vision problems (e.g. ImageNet classification, P ascal VOC object detection as well as CityScapes semantic segmentation) that the proposed method, dubbed SPIQ, achieves accuracies rivalling dynamic approaches with static-level inference speed, significantly outperforming state-of-the-art quantization methods on every benchmark.

**********

Image Segmentation-Based Unsupervised Multiple Objects Discovery

Sandra Kara, Hejer Ammar, Florian Chabot, Quoc-Cuong Pham; Proceedings of the IE EE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 32 77-3286

Unsupervised object discovery aims to localize objects in images, while removing the dependence on annotations required by most deep learning-based methods. To address this problem, we propose a fully unsupervised, bottom-up approach, for m ultiple objects discovery. The proposed approach is a two-stage framework. First , instances of object parts are segmented by using the intra-image similarity be tween self-supervised local features. The second step merges and filters the obj ect parts to form complete object instances. The latter is performed by two CNN models that capture semantic information on objects from the entire dataset. We demonstrate that the pseudo-labels generated by our method provide a better prec ision-recall trade-off than existing single and multiple objects discovery metho ds. In particular, we provide state-of-the-art results for both unsupervised cla ss-agnostic object detection and unsupervised image segmentation.

**********

TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking

Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, Zicheng Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4870-4880

Tracking multiple objects in videos relies on modeling the spatial-temporal inte ractions of the objects. In this paper, we propose TransMOT, which leverages pow erful graph transformers to efficiently model the spatial and temporal interacti ons among the objects. TransMOT is capable of effectively modeling the interacti ons of a large number of objects by arranging the trajectories of the tracked ta rgets and detection candidates as a set of sparse weighted graphs, and construct ing a spatial graph transformer encoder layer, a temporal transformer encoder la yer, and a spatial graph transformer decoder layer based on the graphs. Through end-to-end learning, TransMOT can exploit the spatial-temporal clues to directly estimate association from a large number of loosely filtered detection predicti ons for robust MOT in complex scenes. The proposed method is evaluated on multip le benchmark datasets, including MOT15, MOT16, MOT17, and MOT20, and it achieves state-of-the-art performance on all the datasets.

**********

DSFormer: A Dual-Domain Self-Supervised Transformer for Accelerated Multi-Contra st MRI Reconstruction

Bo Zhou, Neel Dey, Jo Schlemper, Seyed Sadegh Mohseni Salehi, Chi Liu, James S. Duncan, Michal Sofka; Proceedings of the IEEE/CVF Winter Conference on Applicati ons of Computer Vision (WACV), 2023, pp. 4966-4975

Multi-contrast MRI (MC-MRI) captures multiple complementary imaging modalities t o aid in radiological decision-making. Given the need for lowering the time cost of multiple acquisitions, current deep accelerated MRI reconstruction networks focus on exploiting the redundancy between multiple contrasts. However, existing

works are largely supervised with paired data and/or prohibitively expensive fully-sampled MRI sequences. Further, reconstruction networks typically rely on convolutional architectures which are limited in their capacity to model long-range interactions and may lead to suboptimal recovery of fine anatomical detail. To these ends, we present a dual-domain self-supervised transformer (DSFormer) for accelerated MC-MRI reconstruction. DSFormer develops a deep conditional cascade transformer (DCCT) consisting of cascaded Swin transformer reconstruction networks (SwinRN) trained under two deep conditioning strategies to enable MC-MRI information sharing. We further use a dual-domain (image and k-space) self-supervised learning strategy for DCCT to alleviate the costs of acquiring fully sampled training data. DSFormer generates high-fidelity reconstructions which outperform current fully-supervised baselines. Moreover, we find that DSFormer achieves nearly the same performance when trained either with full supervision or with the proposed self-supervision.

**********

Towards Generating Ultra-High Resolution Talking-Face Videos With Lip Synchronization

Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5209-5218

Talking-face video generation works have achieved state-of-the-art results in synthesizing videos with lip synchronization. However, most of the previous works deal with low-resolution talking-face videos (up to 256x256 pixels), thus, generating extremely high-resolution videos still remains a challenge. We take a giant leap in this work and propose a novel method to synthesize talking-face videos at resolutions as high as 4K! Our task presents several key challenges: (i) Scaling the existing methods to such high resolutions is resource-constrained, both in terms of compute and the availability of very high-resolution datasets, (ii) The synthesized videos need to be spatially and temporally coherent. The sheer number of pixels that the model needs to generate while maintaining the temporal consistency at the video level makes this task non-trivial and has never been attempted before in literature. To address these issues, we propose to train the lip-sync generator in a compact Vector Quantized (VQ) space for the first time. Our core idea to encode the faces in a compact 16x16 representation allows us to model high-resolution videos. In our framework, we learn the lip movements in the quantized space on the newly collected 4K Talking Faces (4KTF) dataset. Our approach is speaker agnostic and can handle various languages and voices. We benchmark our technique against several competitive works and show that we can achieve a remarkable 64-times more pixels than the current state-of-the-art! Our supplementary demo video depicts additional qualitative results, comparisons, and several real-world applications, like professional movie editing enabled by our model.

**********

Global-Local Self-Distillation for Visual Representation Learning

Tim Lebailly, Tinne Tuytelaars; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1441-1450

The downstream accuracy of self-supervised methods is tightly linked to the proxy task solved during training and the quality of the gradients extracted from it. Richer and more meaningful gradients updates are key to allow self-supervised methods to learn better and in a more efficient manner. In a typical self-distillation framework, the representation of two augmented images are enforced to be coherent at the global level. Nonetheless, incorporating local cues in the proxy task can be beneficial and improve the model accuracy on downstream tasks. This leads to a dual objective in which, on the one hand, coherence between global-representations is enforced and on the other, coherence between local-representations is enforced. Unfortunately, an exact correspondence mapping between two sets of local-representations does not exist making the task of matching local-representations from one augmentation to another non-trivial. We propose to leverage the spatial information in the input images to obtain geometric matchings and compare this geometric approach against previous methods based on similarity matc

hings. Our study shows that not only 1) geometric matchings perform better than similarity based matchings in low-data regimes but also 2) that similarity based matchings are highly hurtful in low-data regimes compared to the vanilla baseline without local self-distillation. The code is available at https://github.com/tileb1/global-local-self-distillation.

**********

A Morphology Focused Diffusion Probabilistic Model for Synthesis of Histopathology Images

Puria Azadi Moghadam, Sanne Van Dalen, Karina C. Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, Ali Bashashati; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2000-2009

Visual microscopic study of diseased tissue by pathologists has been the cornerstone for cancer diagnosis and prognostication for more than a century. Recently, deep learning methods have made significant advances in the analysis and classification of tissue images. However, there has been limited work on the utility of such models in generating histopathology images. These synthetic images have several applications in pathology including utilities in education, proficiency testing, privacy, and data sharing. Recently, diffusion probabilistic models were introduced to generate high quality images. Here, for the first time, we investigate the potential use of such models along with prioritized morphology weighting and color normalization to synthesize high quality histopathology images of brain cancer. Our detailed results show that diffusion probabilistic models are capable of synthesizing a wide range of histopathology images and have superior performance compared to generative adversarial networks.

**********

STAR-Transformer: A Spatio-Temporal Cross Attention Transformer for Human Action Recognition

Dasom Ahn, Sangwon Kim, Hyunsu Hong, Byoung Chul Ko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3330-3339

In action recognition, although the combination of spatio-temporal videos and skeleton features can improve the recognition performance, a separate model and balancing feature representation for cross-modal data are required. To solve these problems, we propose Spatio-TemporAl cRoss (STAR)-transformer, which can effectively represent two cross-modal features as a recognizable vector. First, from the input video and skeleton sequence, video frames are output as global grid tokens and skeletons are output as joint map tokens, respectively. These tokens are then aggregated into multi-class tokens and input into STAR-transformer. The STAR-transformer encoder consists of a full spatio-temporal attention (FAttn) module and a proposed zigzag spatio-temporal attention (ZAttn) module. Similarly, the continuous decoder consists of a FAttn module and a proposed binary spatio-temporal attention (BAttn) module. STAR-transformer learns an efficient multi-feature representation of the spatio-temporal features by properly arranging pairings of the FAttn, ZAttn, and BAttn modules. Experimental results on the Penn-Action, NTU-RGB+D 60, and 120 datasets show that the proposed method achieves a promising improvement in performance in comparison to previous state-of-the-art methods.

**********

Accelerating Self-Supervised Learning via Efficient Training Strategies

Mustafa Taha Koçyi■it, Timothy M. Hospedales, Hakan Bilen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5654-5664

Recently the focus of the computer vision community has shifted from expensive supervised learning towards self-supervised learning of visual representations. While the performance gap between supervised and self-supervised has been narrowing, the time for training self-supervised deep networks remains an order of magnitude larger than its supervised counterparts, which hinders progress, imposes carbon cost, and limits societal benefits to institutions with substantial resources. Motivated by these issues, this paper investigates reducing the training time of recent self-supervised methods by various model-agnostic strategies that h

ave not been used for this problem. In particular, we study three strategies: an extendable cyclic learning rate schedule, a matching progressive augmentation magnitude and image resolutions schedule, and a hard positive mining strategy based on augmentation difficulty. We show that all three methods combined lead up to 2.7 times speed-up in the training time of several self-supervised methods while retaining comparable performance to the standard self-supervised learning setting.

**********

LAVA: Label-Efficient Visual Learning and Adaptation

Islam Nassar, Munawar Hayat, Ehsan Abbasnejad, Hamid Rezatofighi, Mehrtash Harandi, Gholamreza Haffari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 147-156

We present LAVA, a simple yet effective method for multi-domain visual transfer learning with limited data. LAVA builds on a few recent innovations to enable adapting to partially labelled datasets with class and domain shifts. First, LAVA learns self-supervised visual representations on the source dataset and ground them using class label semantics to overcome transfer collapse problems associated with supervised pretraining. Secondly, LAVA maximises the gains from unlabelled target data via a novel method which uses multi-crop augmentations to obtain highly robust pseudo-labels. By combining these ingredients, LAVA achieves a new state-of-the-art on ImageNet semi-supervised protocol, as well as on 7 out of 10 datasets in multi-domain few-shot learning on the Meta-dataset.

**********

AT-DDPM: Restoring Faces Degraded by Atmospheric Turbulence Using Denoising Diffusion Probabilistic Models

Nithin Gopalakrishnan Nair, Kangfu Mei, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3434-3443

Although many long-range imaging systems are designed to support extended vision applications, a natural obstacle to their operation is degradation due to atmospheric turbulence. Atmospheric turbulence causes significant degradation to image quality by introducing blur and geometric distortion. In recent years, various deep learning-based single image atmospheric turbulence mitigation methods, including CNN-based and GAN inversion-based, have been proposed in the literature which attempt to remove the distortion in the image. However, some of these methods are difficult to train and often fail to reconstruct facial features and produce unrealistic results, especially in the case of high turbulence. Denoising Diffusion Probabilistic Models (DDPMs) have recently gained some traction because of their stable training process and their ability to generate high quality images. In this paper, we propose the first DDPM-based solution for the problem of atmospheric turbulence mitigation. We also propose a fast sampling technique for reducing the inference times for conditional DDPMs. Extensive experiments are conducted on synthetic and real-world data to show the significance of our model. To facilitate further research, all codes and pretrained models are publically available at http://github.com/Nithin-GK/AT-DDPM

**********

Benchmarking Visual Localization for Autonomous Navigation

Lauri Suomela, Jussi Kalliola, Atakan Dag, Harry Edelman, Joni-Kristian Kämäräinen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2945-2955

This work introduces a simulator-based benchmark for visual localization in the autonomous navigation context. The dynamic benchmark enables investigation of how variables such as the time of day, weather, and camera perspective affect the navigation performance of autonomous agents that utilize visual localization for closed-loop control. The experimental part of the paper studies the effects of four such variables by evaluating state-of-the-art visual localization methods as part of the motion planning module of an autonomous navigation stack. The results show major variation in the suitability of the different methods for vision-based navigation. To the authors' best knowledge, the proposed benchmark is the first to study modern visual localization methods as part of a complete navigati

on stack. We make the benchmark available at https://github.com/lasuomela/carla_vloc_benchmark.

**********

Toward Edge-Efficient Dense Predictions With Synergistic Multi-Task Neural Architecture Search

Thanh Vu, Yanqi Zhou, Chunfeng Wen, Yueqi Li, Jan-Michael Frahm; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1400-1410

In this work, we propose a novel and scalable solution to address the challenges of developing efficient dense predictions on edge platforms. Our first key insight is that MultiTask Learning (MTL) and hardware-aware Neural Architecture Search (NAS) can work in synergy to greatly benefit on-device Dense Predictions (DP). Empirical results reveal that the joint learning of the two paradigms is surprisingly effective at improving DP accuracy, achieving superior performance over both the transfer learning of single-task NAS and prior state-of-the-art approaches in MTL, all with just 1/10th of the computation. To the best of our knowledge, our framework, named EDNAS, is the first to successfully leverage the synergistic relationship of NAS and MTL for DP. Our second key insight is that the standard depth training for multi-task DP can cause significant instability and noise to MTL evaluation. Instead, we propose JAReD, an improved, easy-to-adopt Joint Absolute-Relative Depth loss, that reduces up to 88% of the undesired noise while simultaneously boosting accuracy. We conduct extensive evaluations on standard datasets, benchmark against strong baselines and state-of-the-art approaches, as well as provide an analysis of the discovered optimal architectures.

**********

SONGs: Self-Organizing Neural Graphs

■ukasz Struski, Tomasz Danel, Marek ■mieja, Jacek Tabor, Bartosz Zieli■ski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3848-3857

Recent years have seen a surge in research on combining deep neural networks with other methods, including decision trees and graphs. There are at least three advantages of incorporating decision trees and graphs: they are easy to interpret since they are based on sequential decisions, they can make decisions faster, and they provide a hierarchy of classes. However, one of the well-known drawbacks of decision trees, as compared to decision graphs, is that decision trees cannot reuse the decision nodes. Nevertheless, decision graphs were not commonly used in deep learning due to the lack of efficient gradient-based training techniques. In this paper, we fill this gap and provide a general paradigm based on Markov processes, which allows for efficient training of the special type of decision graphs, which we call Self-Organizing Neural Graphs (SONG). We provide a theoretical study on SONG, complemented by experiments conducted on Letter, Connect4, MNIST, CIFAR, and TinyImageNet datasets, showing that our method performs on par or better than existing decision models.

**********

Online Knowledge Distillation for Multi-Task Learning

Geethu Miriam Jacob, Vishal Agarwal, Björn Stenger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2359-2368

Multi-task learning (MTL) has found wide application in computer vision tasks. It uses a common backbone network allowing shared feature computation for different tasks such as semantic segmentation, depth- and normal estimation. In many cases negative transfer, i.e. impaired performance in the target domain, causes the MTL accuracy to be lower than simply training the corresponding single-task networks. To mitigate this issue, we propose an online knowledge distillation method for MTL, where single-task networks are trained simultaneously with the MTL network, guiding the optimization process. We propose selectively training layers for each task using an adaptive feature distillation (AFD) loss with an online task weighting (OTW) scheme. This task-wise feature distillation enables the MTL network to be trained in a similar way to the single-task networks. On the NYUv2 and Cityscapes datasets we show improvements over a baseline MTL model by 6.22% and 9.19%, respectively, and better performance than recent MTL methods. We va

lidate our design choices, including the use of the online task weighting and th
e adaptive feature distillation loss in ablative experiments.
**********

A Simple and Efficient Pipeline To Build an End-to-End Spatial-Temporal Action D
etector
Lin Sui, Chen-Lin Zhang, Lixin Gu, Feng Han; Proceedings of the IEEE/CVF Winter
Conference on Applications of Computer Vision (WACV), 2023, pp. 5999-6008
Spatial-temporal action detection is a vital part of video understanding. Curren
t spatial-temporal action detection methods mostly use an object detector to obt
ain person candidates and classify these person candidates into different action
 categories. So-called two-stage methods are heavy and hard to apply in real-wor
ld applications. Some existing methods build one-stage pipelines, But a large pe
rformance drop exists with the vanilla one-stage pipeline and extra classificati
on modules are needed to achieve comparable performance. In this paper, we explo
re a simple and effective pipeline to build a strong one-stage spatial-temporal
action detector. The pipeline is composed by two parts: one is a simple end-to-e
nd spatial-temporal action detector. The proposed end-to-end detector has minor
architecture changes to current proposal-based detectors and does not add extra
action classification modules. The other part is a novel labeling strategy to ut
ilize unlabeled frames in sparse annotated data. We named our model as SE-STAD.
The proposed SE-STAD achieves around 2% mAP boost and around 80% FLOPs reduction
. Our code will be released at https://github.com/4paradigm-CV/SE-STAD.
**********

Burst Reflection Removal Using Reflection Motion Aggregation Cues
B. H. Pawan Prasad, Green Rosh K. S., Lokesh R. B., Kaushik Mitra; Proceedings o
f the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023
, pp. 239-248
Single image reflection removal has attracted lot of interest in the recent past
 with data driven approaches demonstrating significant improvements. However dee
p learning based approaches for multi-image reflection removal remains relativel
y less explored. The existing multi-image methods require input images to be cap
tured at sufficiently different view points with wide baselines. This makes it c
umbersome for the user who is required to capture the scene by moving the camera
 in multiple directions. A more convenient way is to capture a burst of images i
n a short time duration without providing any specific instructions to the user.
 A burst of images captured on a hand-held device provide crucial cues that rely
 on the subtle handshakes created during the capture process to separate the ref
lection and the transmission layers. In this paper, we propose a multi-stage dee
p learning based approach for burst reflection removal. In the first stage, we p
erform reflection suppression on the individual images. In the second stage, a n
ovel reflection motion aggregation (RMA) cue is extracted that emphasizes the tr
ansmission layer more than the reflection layer to aid better layer separation.
In our final stage we use this RMA cue as a guide to remove reflections from the
 input. We provide the first real world burst images dataset along with ground t
ruth for reflection removal that can enable future benchmarking. We evaluate bot
h qualitatively and quantitatively to demonstrate the superiority of the propose
d approach. Our method achieves  2 dB improvement in PSNR over single image base
d methods and  1 dB over multi-image based methods.
**********

Rethinking the Data Annotation Process for Multi-View 3D Pose Estimation With Ac
tive Learning and Self-Training
Qi Feng, Kun He, He Wen, Cem Keskin, Yuting Ye; Proceedings of the IEEE/CVF Wint
er Conference on Applications of Computer Vision (WACV), 2023, pp. 5695-5704
Pose estimation of the human body and hands is a fundamental problem in computer
 vision, and learning-based solutions require a large amount of annotated data.
In this work, we improve the efficiency of the data annotation process for 3D po
se estimation problems with Active Learning (AL) in a multi-view setting. AL sel
ects examples with the highest value to annotate under limited annotation budget
s (time and cost), but choosing the selection strategy is often nontrivial. We p
resent a framework to efficiently extend existing single-view AL strategies. We

then propose two novel AL strategies that make full use of multi-view geometry. Moreover, we demonstrate additional performance gains by incorporating pseudo-la bels computed during the AL process, which is a form of self-training. Our syste m significantly outperforms simulated annotation baselines in 3D body and hand p ose estimation on two large-scale benchmarks: CMU Panoptic Studio and InterHand2 .6M. Notably, on CMU Panoptic Studio, we are able to reduce the turn-around time by 60% and annotation cost by 80% when compared to the conventional annotation process.

**********

Context-Empowered Visual Attention Prediction in Pedestrian Scenarios

Igor Vozniak, Philipp Müller, Lorena Hell, Nils Lipp, Ahmed Abouelazm, Christian Müller; Proceedings of the IEEE/CVF Winter Conference on Applications of Comput er Vision (WACV), 2023, pp. 950-960

Effective and flexible allocation of visual attention is key for pedestrians who have to navigate to a desired goal under different conditions of urgency and sa fety preferences. While automatic modelling of pedestrian attention holds great promise to improve simulations of pedestrian behavior, current saliency predicti on approaches mostly focus on generic free-viewing scenarios and do not reflect the specific challenges present in pedestrian attention prediction. In this pape r, we present Context-SalNET, a novel encoder-decoder architecture that explicit ly addresses three key challenges of visual attention prediction in pedestrians: First, Context-SalNET explicitly models the context factors urgency and safety preference in the latent space of the encoder-decoder model. Second, we propose the exponentially weighted mean squared error loss (ew-MSE) that is able to bett er cope with the fact that only a small part of the ground truth saliency maps c onsist of non-zero entries. Third, we explicitly model epistemic uncertainty to account for the fact that training data for pedestrian attention prediction is l imited. To evaluate Context-SalNET, we recorded the first dataset of pedestrian visual attention in VR that includes explicit variation of the context factors u rgency and safety preference. Context-SalNET achieves clear improvements over st ate-of-the-art saliency prediction approaches as well as over ablations. Our nov el dataset will be made fully available and can serve as a valuable resource for further research on pedestrian attention prediction.

**********

COPE: End-to-End Trainable Constant Runtime Object Pose Estimation

Stefan Thalhammer, Timothy Patten, Markus Vincze; Proceedings of the IEEE/CVF Wi nter Conference on Applications of Computer Vision (WACV), 2023, pp. 2860-2870

State-of-the-art object pose estimation handles multiple instances in a test ima ge by using multi-model formulations: detection as a first stage and then separa tely trained networks per object for 2D-3D geometric correspondence prediction a s a second stage. Poses are subsequently estimated using the Perspective-n-Point s algorithm at runtime. Unfortunately, multi-model formulations are slow and do not scale well with the number of object instances involved. Recent approaches s how that direct 6D object pose estimation is feasible when derived from the afor ementioned geometric correspondences. We present an approach that learns an inte rmediate geometric representation of multiple objects to directly regress 6D pos es of all instances in a test image. The inherent end-to-end trainability overco mes the requirement of separately processing individual object instances. By cal culating the mutual Intersection-over-Unions, pose hypotheses are clustered into distinct instances, which achieves negligible runtime overhead with respect to the number of object instances. Results on multiple challenging standard dataset s show that the pose estimation performance is superior to single-model state-of -the-art approaches despite being more than 35 times faster. We additionally pr ovide an analysis showing real-time applicability (>24 fps) for images where mor e than 90 object instances are present. Further results show the advantage of su pervising geometric correspondence-based object pose estimation with the 6D pose .

**********

DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network

Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, Ig-

Jae Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Comput
er Vision (WACV), 2023, pp. 5541-5550
Unsupervised approaches for video anomaly detection may not perform as good as s
upervised approaches. However, learning unknown types of anomalies using an unsu
pervised approach is more practical than a supervised approach as annotation is
an extra burden. In this paper, we use isolation tree-based unsupervised cluster
ing to partition the deep feature space of the video segments. The RGB- stream g
enerates a pseudo anomaly score and the flow stream generates a pseudo dynamicit
y score of a video segment. These scores are then fused using a majority voting
scheme to generate preliminary bags of positive and negative segments. However,
these bags may not be accurate as the scores are generated only using the curren
t segment which does not represent the global behavior of a typical anomalous ev
ent. We then use a refinement strategy based on a cross-branch feed-forward netw
ork designed using a popular I3D network to refine both scores. The bags are the
n refined through a segment re-mapping strategy. The intuition of adding the dyn
amicity score of a segment with the anomaly score is to enhance the quality of t
he evidence. The method has been evaluated on three popular video anomaly datase
ts, i.e., UCF-Crime, CCTV-Fights, and UBI-Fights. Experimental results reveal th
at the proposed framework achieves competitive accuracy as compared to the state
-of-the-art video anomaly detection methods.
**********

Exploiting Long-Term Dependencies for Generating Dynamic Scene Graphs
Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, Subarna Tripathi;
Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision
 (WACV), 2023, pp. 5130-5139
Dynamic scene graph generation from a video is challenging due to the temporal d
ynamics of the scene and the inherent temporal fluctuations of predictions. We h
ypothesize that capturing long-term temporal dependencies is the key to effectiv
e generation of dynamic scene graphs. We propose to learn the long-term dependen
cies in a video by capturing the object-level consistency and inter-object relat
ionship dynamics over object-level long-term tracklets using transformers. Exper
imental results demonstrate that our "Dynamic Scene Graph Detection Transformer"
 (DSG-DETR) outperforms state-of-the-art methods by a significant margin on the
benchmark dataset Action Genome. Our ablation studies validate the effectiveness
 of each component of the proposed approach. The source code is available at htt
ps://github.com/Shengyu-Feng/DSG-DETR.
**********

Two-Level Data Augmentation for Calibrated Multi-View Detection
Martin Engilberge, Haixin Shi, Zhiye Wang, Pascal Fua; Proceedings of the IEEE/C
VF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 128-13
6
Data augmentation has proven its usefulness to improve model generalization and
performance. While it is commonly applied in computer vision application when it
 comes to multi-view systems, it is rarely used. Indeed geometric data augmentat
ion can break the alignment among views. This is problematic since multi-view da
ta tend to be scarce and it is expensive to annotate. In this work we propose to
 solve this issue by introducing a new multi-view data augmentation pipeline tha
t preserves alignment among views. Additionally to traditional augmentation of t
he input image we also propose a second level of augmentation applied directly a
t the scene level. When combined with our simple multi-view detection model, our
 two-level augmentation pipeline outperforms all existing baselines by a signifi
cant margin on the two main multi-view multi-person detection datasets WILDTRACK
 and MultiviewX.
**********

Multi-Frame Attention With Feature-Level Warping for Drone Crowd Tracking
Takanori Asanomi, Kazuya Nishimura, Ryoma Bise; Proceedings of the IEEE/CVF Wint
er Conference on Applications of Computer Vision (WACV), 2023, pp. 1664-1673
Drone crowd tracking has various applications such as crowd management and video
 surveillance. Unlike in general multi-object tracking, the size of the objects
to be tracked are small, and the ground truth is given by a point-level annotati

on, which has no region information. This causes the lack of discriminative features for finding the same objects from many similar objects. Thus, similarity-based trackingtechniques, which are widely used for multi-object tracking with bounding-box, are difficult to use. To deal with this problem, we take into account the temporal context of the local area. To aggregate temporal context in a local area, we propose a multi-frame attention with feature-level warping. The feature-level warping can align the features of the same object in multiple frame, and then multi-frame attention can effectively aggregate the temporal context from the warped features. The experimental results show the effectiveness of our method. Our method outperformed the state-of-the-art method in DroneCrowd dataset.
**********

MonoEdge: Monocular 3D Object Detection Using Local Perspectives
Minghan Zhu, Lingting Ge, Panqu Wang, Huei Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 643-652
We propose a novel approach for monocular 3D object detection by leveraging local perspective effects of each object. While the global perspective effect shown as size and position variations has been exploited for monocular 3D detection extensively, the local perspectives has long been overlooked. We propose a new regression target named keyedge-ratios as the parameterization of the local shape distortion to account for the local perspective, and derive the object depth and yaw angle from it. Theoretically, this approach does not rely on the absolute size or position of the objects in the image, therefore independent of the camera intrinsic parameters. This approach provides a new perspective for monocular 3D reasoning and can be plugged in flexibly to existing monocular 3D object detection frameworks. We demonstrate effectiveness and superior performance over strong baseline methods in multiple datasets.
**********

GAFNet: A Global Fourier Self Attention Based Novel Network for Multi-Modal Downstream Tasks
Onkar Susladkar, Gayatri Deshmukh, Dhruv Makwana, Sparsh Mittal, R. Sai Chandra Teja, Rekha Singhal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5242-5251
In "vision and language" problems, multimodal inputs are simultaneously processed for combined visual and textual understanding for image-text embedding. In this paper, we discuss the necessity of considering the difference between the feature space and the distribution when performing multimodal learning. We deal with this problem through deep learning and a generative model approach. We introduce a novel network, GAFNet (Global Attention Fourier Net) which learns through large-scale pre-training over three image-text datasets (COCO, SBU, and CC-3M), for achieving high performance on downstream vision and language tasks. We propose a GAF (Global Attention Fourier) module, which integrates multiple modalities into one latent space. GAF module is independent of the type of modality and it allows combining shared representations at each stage. There are various ways of thinking about the relationships between different modalities, which directly affect the model's design. Global attention is not considered as in conventional multimodal learning. A GAF-based model can work for any modality (language, image, audio, category) and is designed to be used for different tasks. In contrast to previous research, our work considers visual grounding as a pretrainable and transferable quality instead of something that must be trained from scratch. Experimental results demonstrate that our technique is competitive and achieves state-of-the-art performance on a variety of popular downstream vision-language tasks, including image generation and image-text retrieval.
**********

Sim2RealVS: A New Benchmark for Video Stabilization With a Strong Baseline
Qi Rao, Xin Yu, Shant Navasardyan, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5406-5415
Video stabilization is highly desirable when videos undergo severe jittering artifacts. The difficulty of obtaining sufficient training data obstructs the development of video stabilization. In this work, we address this issue by presenting a Sim2RealVS benchmark with more than 1,300 pairs of shaky and stable videos. O

ur benchmark is curated by an in-game simulator with diverse scenes and various jittering effects. Moreover, we propose a simple yet strong baseline approach, named Motion-Trajectory Smoothing Network (MTSNet), by fully exploiting our Sim2RealVS data. Our MTSNet consists of three main steps: motion estimation, global trajectory smoothing and frame warping. In motion estimation, we design a Motion Correction and Completion (MCC) module to rectify the optical flow with low confidence, such as in textureless regions, thus providing more consistent motion estimation for next steps. Benefiting from our synthetic data, we can explicitly learn a Trajectory Smoothing Transformer (TST) with ground-truth supervision to smooth global trajectories. In training TST, we propose two fully-supervised losses, i.e., a motion magnitude similarity loss and a motion tendency similarity loss. After training, our TST is able to produce smooth motion trajectories for the shaky input videos. Extensive qualitative and quantitative results demonstrate that our MTSNet achieves superior performance on both synthetic and real-world data.

**********

Domain Adaptation Using Self-Training With Mixup for One-Stage Object Detection

Jitender Maurya, Keyur R. Ranipa, Osamu Yamaguchi, Tomoyuki Shibata, Daisuke Kobayashi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4189-4198

In this paper, we present an end-to-end domain adaptation technique that utilizes both feature distribution alignment and Self-Training effectively for object detection. One set of methods for domain adaptation relies on feature distribution alignment and adapts models on an unlabeled target domain by learning domain invariant representations through adversarial loss. Although this approach is effective, it may not be adequate or even have an adverse effect when domain shifts are large and inconsistent. Another set of methods utilizes Self-Training which relies on pseudo labels to approximate the target domain distribution directly. However, it can also have a negative impact on the model performance due to erroneous pseudo labels. To overcome these two issues, we propose to generate reliable pseudo labels through feature distribution alignment and data distillation. Further, to minimize the adverse effect of incorrect pseudo labels during Self-Training we employ interpolation-based consistency regularization called mixup. While distribution alignment helps in generating more accurate pseudo labels, mixup regularization of Self-Training reduces the adverse effect of less accurate pseudo labels. Both approaches supplement each other and achieve effective adaptation on the target domain which we demonstrate through extensive experiments on one-stage object detector. Experiment results show that our approach achieves a significant performance improvement on multiple benchmark datasets.

**********

Automated Detection of Label Errors in Semantic Segmentation Datasets via Deep Learning and Uncertainty Quantification

Matthias Rottmann, Marco Reese; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3214-3223

In this work, we for the first time present a method for detecting labeling errors in image datasets with semantic segmentation, i.e., pixel-wise class labels. Annotation acquisition for semantic segmentation datasets is time-consuming and requires plenty of human labor. In particular, review processes are time consuming and label errors can easily be overlooked by humans. The consequences are biased benchmarks and in extreme cases also performance degradation of deep neural networks (DNNs) trained on such datasets. DNNs for semantic segmentation yield pixel-wise predictions, which makes detection of labeling errors via uncertainty quantification a complex task. Uncertainty is particularly pronounced at the transitions between connected components of the prediction. By lifting the consideration of uncertainty to the level of predicted components, we enable the usage of DNNs together with component-level uncertainty quantification for the detection of labeling errors. We present a principled approach to benchmarking the task of label error detection by dropping labels from the Cityscapes dataset as well from a dataset extracted from the CARLA driving simulator, where in the latter case we have the labels under control. Our experiments show that our approach is

able to detect the vast majority of labeling errors while controlling the number of false label error detections. Furthermore, we apply our method to semantic segmentation datasets frequently used by the computer vision community and present a collection of labeling errors along with sample statistics.
**********

Joint Video Rolling Shutter Correction and Super-Resolution
Akash Gupta, Sudhir Kumar Singh, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4946-4955
With the prevalence of CMOS cameras in many computer vision applications, there is an increase in the appearance of rolling shutter (RS) artifacts in captured videos. However, existing video super-resolution algorithms assume that the motion is globally consistent in each video frame and no rolling shutter effect is present. The problem of video super-resolution for video captured using RS cameras is challenging as the model needs to learn the row-wise local pixel displacements and the global structure of the frame for RS correction and super-resolution, simultaneously. Different from existing works, we address a more realistic problem of joint rolling shutter correction and super-resolution (RS-SR). We introduce a novel architecture, deformable Patch Attention Network (PatchNet), that utilizes patch-recurrence property along with deformable receptive fields to learn the global and local structure of the video. Specifically, PatchNet leverages bi-directional motion field in the feature space to extract relevant information from neighboring patches using attention mechanism, and deformable fields using deformable convolutions to extract local pixel-level information for joint rolling shutter correction and super-resolution. Our work is the first to tackle the task of RS correction and super-resolution on the recently released BS-RSCD dataset. Experiments on the BS-RSCD dataset and FastecRS datasets demonstrate that our model performs favorably against various state-of-the-art approaches.
**********

Self-Supervised Correspondence Estimation via Multiview Registration
Mohamed El Banani, Ignacio Rocco, David Novotny, Andrea Vedaldi, Natalia Neverova, Justin Johnson, Ben Graham; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1216-1225
Video provides us with the spatio-temporal consistency needed for visual learning. Recent approaches have utilized this signal to learn correspondence estimation from close-by frame pairs. However, by only relying on close-by frame pairs, those approaches miss out on the richer long-range consistency between more distant overlapping frames. To address this, we propose a self-supervised approach for correspondence estimation that learns from multiview consistency in short RGB-D video sequences. Our approach combines pairwise correspondence estimation and registration with a novel SE(3) transformation synchronization algorithm. Our key insight is that self-supervised multiview registration allows us to obtain correspondences over longer time frames, which increases both the diversity and difficulty of sampled pairs. We evaluate our approach on indoor scenes for correspondence estimation and RGB-D pointcloud registration and find that we can perform on-par with prior supervised approaches.
**********

Anisotropic Multi-Scale Graph Convolutional Network for Dense Shape Correspondence
Mohammad Farazi, Wenhui Zhu, Zhangsihao Yang, Yalin Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3146-3155
This paper studies 3D dense shape correspondence, a key shape analysis application in computer vision and graphics. We introduce a novel hybrid geometric deep learning-based model that learns geometrically meaningful and discretization-independent features. The proposed framework has a U-Net model as the primary node feature extractor, followed by a successive spectral-based graph convolutional network. To create a diverse set of filters, we use anisotropic wavelet basis filters, being sensitive to both different directions and band-passes. This filter set overcomes the common over-smoothing behavior of conventional graph neural net

works. To further improve the model's performance, we add a function that perturbs the feature maps in the last layer ahead of fully connected layers, forcing the network to learn more discriminative features overall. The resulting correspondence maps show state-of-the-art performance on the benchmark datasets based on average geodesic errors and superior robustness to discretization in 3D meshes. Our approach provides new insights and practical solutions to the dense shape correspondence research.

**********

Recipe2Video: Synthesizing Personalized Videos From Recipe Texts

Prateksha Udhayanan, Suryateja BV, Parth Laturia, Dev Chauhan, Darshan Khandelwal, Stefano Petrangeli, Balaji Vasan Srinivasan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2268-2277

Procedural texts are a special type of documents that contain complex textual descriptions for carrying out a sequence of instructions. Due to the lack of visual cues, it often becomes difficult to consume the textual information effectively. In this paper, we focus on recipes - a particular type of procedural document and introduce a novel deep-learning driven system - Recipe2Video that automatically converts a recipe document into a multimodal illustrative video. Our method employs novel retrieval and re-ranking methods to select the best set of images and videos that can provide the desired illustration. We formulate a Viterbi-based optimization algorithm to stitch together a coherent video that combines the visual cues, text and voice-over to present an enhanced mode of consumption. We design automated metrics and compare performance across several baselines on two recipe datasets (RecipeQA, Tasty Videos). Our results on downstream tasks and human studies indicate that Recipe2Video captures the semantic and sequential information of the input in the generated video.

**********

Exemplar Guided Deep Neural Network for Spatial Transcriptomics Analysis of Gene Expression Prediction

Yan Yang, Md Zakir Hossain, Eric A. Stone, Shafin Rahman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5039-5048

Spatial transcriptomics (ST) is essential for understanding diseases and developing novel treatments. It measures gene expression of each fine-grained area (i.e., different windows) in the tissue slide with low throughput. This paper proposes an Exemplar Guided Network (EGN) to accurately and efficiently predict gene expression directly from each window of a tissue slide image. We apply exemplar learning to dynamically boost gene expression prediction from nearest/similar exemplars of a given tissue slide image window. Our EGN framework composes of three main components: 1) an extractor to structure a representation space for unsupervised exemplar retrievals; 2) a vision transformer (ViT) backbone to progressively extract representations of the input window; and 3) an Exemplar Bridging (EB) block to adaptively revise the intermediate ViT representations by using the nearest exemplars. Finally, we complete the gene expression prediction task with a simple attention-based prediction block. Experiments on standard benchmark datasets indicate the superiority of our approach when comparing with the past state-of-the-art (SOTA) methods.

**********

The Fully Convolutional Transformer for Medical Image Segmentation

Athanasios Tragakis, Chaitanya Kaul, Roderick Murray-Smith, Dirk Husmeier; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3660-3669

We propose a novel transformer model, capable of segmenting medical images of varying modalities. Challenges posed by the fine-grained nature of medical image analysis mean that the adaptation of the transformer for their analysis is still at nascent stages. The overwhelming success of the UNet lay in its ability to appreciate the fine-grained nature of the segmentation task, an ability which existing transformer based models do not currently posses. To address this shortcoming, we propose The Fully Convolutional Transformer (FCT), which builds on the proven ability of Convolutional Neural Networks to learn effective image represent

ations, and combines them with the ability of Transformers to effectively captur e long-term dependencies in its inputs. The FCT is the first fully convolutional Transformer model in medical imaging literature. It processes its input in two stages, where first, it learns to extract long range semantic dependencies from the input image, and then learns to capture hierarchical global attributes from the features. FCT is compact, accurate and robust. Our results show that it outp erforms all existing transformer architectures by large margins across multiple medical image segmentation datasets of varying data modalities without the need for any pre-training. FCT outperforms its immediate competitor on the ACDC datas et by 1.3%, on the Synapse dataset by 4.4%, on the Spleen dataset by 1.2% and on ISIC 2017 dataset by 1.1% on the dice metric, with up to five times fewer param eters. On the ACDC Post-2017-MICCAI-Challenge online test set, our model sets a new state-of-the-art on unseen MRI test cases outperforming large ensemble model s as well as nnUNet with considerably fewer parameters. Our code, environments a nd models will be available via GitHub.

**********

CameraPose: Weakly-Supervised Monocular 3D Human Pose Estimation by Leveraging I n-the-Wild 2D Annotations

Cheng-Yen Yang, Jiajia Luo, Lu Xia, Yuyin Sun, Nan Qiao, Ke Zhang, Zhongyu Jiang , Jenq-Neng Hwang, Cheng-Hao Kuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 2924-2933

To improve the generalization of 3D human pose estimators, many existing deep le arning based models focus on adding different augmentations to training poses. H owever, data augmentation techniques are limited to the "seen" pose combinations and hard to infer poses with rare "unseen" joint positions. To address this pro blem, we present CameraPose, a weakly-supervised framework for 3D human pose est imation from a single image, which can not only be applied on 2D-3D pose pairs b ut also on 2D alone annotations. By adding a camera parameter branch, any in-the -wild 2D annotations can be fed into our pipeline to boost the training diversit y and the 3D poses can be implicitly learned by reprojecting back to 2D. Moreove r, CameraPose introduces a refinement network module with confidence-guided loss to further improve the quality of noisy 2D keypoints extracted by 2D pose estim ators. Experimental results demonstrate that the CameraPose brings in clear impr ovements on cross-scenario datasets. Notably, it outperforms the baseline method by 3mm on the most challenging dataset 3DPW. In addition, by combining our prop osed refinement network module with existing 3D pose estimators, their performan ce can be improved in cross-scenario evaluation.

**********

BirdSoundsDenoising: Deep Visual Audio Denoising for Bird Sounds

Youshan Zhang, Jialu Li; Proceedings of the IEEE/CVF Winter Conference on Applic ations of Computer Vision (WACV), 2023, pp. 2248-2257

Audio denoising has been explored for decades using both traditional and deep le arning-based methods. However, these methods are still limited to either manuall y added artificial noise or lower denoised audio quality. To overcome these chal lenges, we collect a large-scale natural noise bird sound dataset. We are the fi rst to transfer the audio denoising problem into an image segmentation problem a nd propose a deep visual audio denoising (DVAD) model. With a total of 14,120 au dio images, we develop an audio ImageMask tool and propose to use a few-shot gen eralization strategy to label these images. Extensive experimental results demon strate that the proposed model achieves state-of-the-art performance. We also sh ow that our method can be easily generalized to speech denoising, audio separati on, audio enhancement, and noise estimation.

**********

Single Stage Weakly Supervised Semantic Segmentation of Complex Scenes

Peri Akiva, Kristin Dana; Proceedings of the IEEE/CVF Winter Conference on Appli cations of Computer Vision (WACV), 2023, pp. 5954-5965

The costly process of obtaining semantic segmentation labels has driven research towards weakly supervised semantic segmentation (WSSS) methods, using only imag e-level, point, or box labels. Such annotations introduce limitations and challe nges that results in overly-tuned methods specialized in specific domains or sce

ne types. The over reliance of image-level based methods on generation of high q
uality class activation maps (CAMs) results in limited applicable dataset comple
xity range, mostly focusing on object centric scenes. Additionally, the lack of
dense annotations requires methods to increase network complexity to obtain addi
tional semantic information, often done through multiple stages of training and
refinement. Here, we present a single-stage approach generalizable to a wide ran
ge of dataset complexities, that is trainable from scratch, without any dependen
cy on pre-trained backbones, classification, or separate refinement tasks. We ut
ilize point annotations to generate reliable, on-the-fly pseudo-masks through re
fined and spatially filtered features. We are to demonstrate SOTA performance on
 benchmark datasets (PascalVOC 2012), as well as significantly outperform other
SOTA WSSS methods on recent real-world datasets (CRAID, CityPersons, IAD, ADE20K
, CityScapes) with up to 28.1% and 22.6% performance boosts compared to our sing
le-stage and multi-stage baselines respectively.
**********

Cross-Resolution Flow Propagation for Foveated Video Super-Resolution
Eugene Lee, Lien-Feng Hsu, Evan Chen, Chen-Yi Lee; Proceedings of the IEEE/CVF W
inter Conference on Applications of Computer Vision (WACV), 2023, pp. 1766-1775
The demand of high-resolution video contents has grown over the years. However,
the delivery of high-resolution video is constrained by either computational res
ources required for rendering or network bandwidth for remote transmission. To r
emedy this limitation, we leverage the eye trackers found alongside existing aug
mented and virtual reality headsets. We propose the application of video super-r
esolution (VSR) technique to fuse low-resolution context with regional high-reso
lution context for resource-constrained consumption of high-resolution content w
ithout perceivable drop in quality. Eye trackers provide us the gaze direction o
f a user, aiding us in the extraction of the regional high-resolution context. A
s only pixels that falls within the gaze region can be resolved by the human eye
, a large amount of the delivered content is redundant as we can't perceive the
difference in quality of the region beyond the observed region. To generate a vi
sually pleasing frame from the fusion of high-resolution region and low-resoluti
on region, we study the capability of a deep neural network of transferring the
context of the observed region to other regions (low-resolution) of the current
and future frames. We label this task a Foveated Video Super-Resolution (FVSR),
as we need to super-resolve the low-resolution regions of current and future fra
mes through the fusion of pixels from the gaze region. We propose Cross-Resoluti
on Flow Propagation (CRFP) for FVSR. We train and evaluate CRFP on REDS dataset
on the task of 8 times FVSR, i.e. a combination of 8 times VSR and the fusion of
 foveated region. Departing from the conventional evaluation of per frame qualit
y using SSIM or PSNR, we propose the evaluation of past foveated region, measuri
ng the capability of a model to leverage the noise present in eye trackers durin
g FVSR.
**********